1050

Scientific Thought
1900–1960

Scientific Thought 1900–1960

A Selective Survey

Edited by
R. HARRÉ
Fellow of Linacre College, Oxford

CLARENDON PRESS · OXFORD
1969

Oxford University Press, Ely House, London W.1
GLASGOW NEW YORK TORONTO MELBOURNE WELLINGTON
CAPE TOWN SALISBURY IBADAN NAIROBI LUSAKA ADDIS ABABA
BOMBAY CALCUTTA MADRAS KARACHI LAHORE DACCA
KUALA LUMPUR SINGAPORE HONG KONG TOKYO

PRINTED IN GREAT BRITAIN
BY SPOTTISWOODE BALLANTYNE AND COMPANY LIMITED
LONDON AND COLCHESTER

Preface

In preparing this collection of essays the aim has been to capture something of the main currents of scientific thought since the beginning of this century. Only the use of a principle of selection could make such an enterprise possible. Each author has been asked to look at the recent history of his subject with an eye to that conceptual novelty which strikes him as having been the most fruitful and the most influential in recent years. Practical applications and development of experimental techniques have had to be treated peripherally, except where they were themselves the fructifying agent. The result is, it is hoped, a fairly complete slice of intellectual history, even though much of the detailed application of the central concepts that the authors have identified has had to be omitted.

Very little of the nineteenth-century picture of the world remains today. A great revolution in concepts and ideas of nature has taken place. The enormous extent of this revolution can be seen in every one of the essays in this book. And yet in each of the essays it is also possible to discern a continuous development of ideas from those of the older view of the world. In many ways the science of the twentieth century has realized the hopes of the past. But just as striking as the revolution in concepts is the identity of method. Mathematical description and analysis, and the invention of models and hypothetical mechanisms, have remained the key to the scientific method. Except perhaps in the science of animal behaviour (ethology), nothing comparable to the revolution of the sixteenth and seventeenth centuries has taken place.

The essays have been assembled in the 'traditional' order of subjects, beginning with the most mathematical, continuing with physics, geology, and chemistry. The biological sciences that follow have been arranged in an order roughly determined by how far their subject-matter can be studied in isolation from actual organisms. The degree of linkage that emerges between fields is remarkable, and with the subjects arranged in the traditional way this continuity can, as it happens, be very strikingly observed in the chapters of this book. If one had to say what was most characteristic of twentieth-century science, as our authors have presented it, it is surely the extraordinary integration of the traditionally independent fields of physics, chemistry, and biology. The work of Bohr in physics joins with the ideas of Lewis in chemistry; the techniques of X-ray crystallography are the key to the deepest problems of molecular biology, and in the end of genetics; the notions of enzymes as organic catalysts, of hormones as messengers, of the chemical basis of electrical conduction in nerves, have led to the gradual unravelling of the mechanisms of life, and promise to provide an explanatory basis for the discoveries of the new science of ethology.

Each author has been asked to try to accomplish two exegetical tasks. The first was to write for other scientists who might want to get an idea of the broad sweep of progress of ideas in scientific subjects in fields remote from their own. The second task was so to present his material that something of the growth of twentieth-century science might be made available to anyone who took the trouble to read this book. Readers of both these classes may be stimulated to read further. So, working on the principle that original sources are usually best, a bibliography of key original papers and books has been included with each chapter. Reference to these bibliographies is included in the author and subject indexes at the end of the book.

Finally I would like to emphasize that each chapter has been written from a personal point of view. Our contributors were asked 'How does science in your area since 1900 strike you?' Other authors whose interests differed might see the relative importance of innovations differently. But science is something done by scientists, and if we have captured something of how the last sixty years looks to those who learned and worked in the period then we have succeeded in the task we have set ourselves.

Linacre College, Oxford R. H.
December 1968

Contents

List of Plates

(*between pp. 56-57*)

1 *Logic*

In this century logic, the analysis of the principles of reasoning, has been transformed by the development and application of the new methods of mathematical logic. We have chosen to illustrate this subject, so to speak, from two extremes. The first sub-section below, due to R. O. Gandy, looks at one of the most important specific contributions, the analysis of the concept of computability, which ultimately led to the development of modern computers. The second, due to G. Kreisel, lists the main lines of development during the first half of this century.

(a) The Concept of Computability

The notion of a rule is fundamental to all rational thought; familiar examples are rules for computation and rules of proof. Rules have two pregnant characteristics: they can be communicated with precision and they are of general application. Anyone can learn to add, and when he has done so he can do sums he has never seen before. Everyone recognizes the importance of the concept of computability for the foundations of mathematics; but this concept has not yet acquired the status in philosophy that it merits. 'Computability' is at least as significant for epistemology as 'analyticity' or 'verifiability'. What is more, and what is surprising, is that an exact definition is available. Surprising because it would be natural to say 'Of course, it all depends on what you mean by "rule"'. The arguments which show that this is not so—that the notion is unambiguous—form a paradigm of philosophical analysis.†

The essential features of this analysis are as follows. (i) The rule is presented by a finite table of instructions (or programme). (ii) The necessary record of the progress of the computation is finite. (iii) At each stage it is only necessary to take cognizance of the current instruction and a limited portion of the record; this knowledge determines uniquely (via the programme) what action is to be taken, what instruction shall next be obeyed, and which portion of the record shall next be consulted. Note that a rule may be applied to an infinite variety of cases, and so no bound can be given to the length of *computations* and their records. But because at each step only a limited portion (e.g. a single symbol) of the record is relevant, it is possible to specify the *rule* by a fixed table of instructions. It should also be stressed that rules may be wholly non-numerical.

The circumstances in which this analysis first appeared are of considerable interest to the historian of ideas. In 1936 Church, Post, and Turing simultaneously and independently published seemingly different, but actually equivalent, definitions of 'computable'. In theory such a definition could have been given at any time after the introduction of algebraic notations; the possibility was indeed recognized by Leibniz. And the general notion of numerical rule underlay Babbage's proposals for an 'analytic engine'. On the other hand, the papers mentioned above preceded by several years the development of electronic digital computers, a development that would in any case have provoked the search for a general definition of 'computable'. In the event, Turing's ideas, and Turing himself, played an important part in the logical design of computers. For the proximate cause of the convergence of ideas in 1936 we must look, not to the theory of computation, but to the researches into mathematical proof theory of Hilbert and his school.‡

We shall now illustrate the application of the concept of computability to epistemology. First, it enables us to characterize once and for all the basic *particular* facts of mathematics; namely, these can all be given the form: such and such a computation has such and such a result. Thus, despite its common use as a stereotype, '$2+2=4$' is not a misleading example. Another example is 'X is a correct formal proof of Y'.

† See, for example, the section entitled 'Church's Thesis' in S. C. Kleene's *Introduction to metamathematics*; or the interested reader may consult the original papers: these have been collected together in *The undecidable*, edited by Martin Davis.

‡ Gödel and Herbrand introduced a notion based on computations *within* a formal system. This notion was given definitive form by Kleene, who showed that it was equivalent to Turing's 'computable'.

Secondly, let us consider theories that depend on a notion of simplicity of pattern. One such is the Pythagorean theory of aesthetics: it is simple patterns that we find harmonious and pleasing.† Others are concerned with assessing the acceptability (or degree of confirmation) of general laws according to their simplicity: a simple law is more improbable than a complex one because, for example, it has fewer parameters; therefore, if both fit a given set of data, the simpler law is to be preferred. Applications of such theories as these usually depend on *ad hoc* notions of simplicity; for example, the degree of a polynomial used in curve-fitting. But the patterns or laws considered are always computable; therefore it is possible to introduce a uniform measure of simplicity, namely, the length of the corresponding programme. Of course, here too there is an arbitrary element: the particular 'programme language' in which the rule of computation is described. But once the general nature of the computation has been decided upon, there will not be much doubt about what shall be the basic elements of the programme language. In particular, it is now becoming clear that the brain organizes its computations quite differently from electronic machines. But when we have a better knowledge of this cerebral organization we may hope to explain in *appropriate* computational terms why certain patterns are more easily apprehended than others. Even crude ideas about the way in which the brain computes suggest reasons why some patterns of linguistic structure are preferred to others that seem at first sight equally convenient.

Finally, once an exact definition of computability is available it is possible to construct problems that cannot be solved by *any* rule; this was indeed one of the aims of the papers published in 1936. What is more, some of these 'unsolvable' problems are of great mathematical interest and had—perforce—defied attempts to find systematic methods of solution. In particular, *there is no rule that will always correctly decide whether a given statement is a consequence of the basic laws of logic.* Some idea of what such unsolvable problems are like can be obtained as follows. Suppose we are given a property P and a (mechanical) method of testing, for each number n, whether or not n has the property P. We now ask: is there *any* number that does have P? The simple-minded way of settling this is to try $n = 0,1,2, \ldots$, in turn. But if no number has P this process will not terminate, and the question will remain unanswered. Of course there may be other ways—'short cuts'—of arriving at a negative answer. But the constructions mentioned above guarantee that

† This theory embraces the golden ratio: for what could be simpler than the continued fraction $1 + 1/1 + 1/1 + \ldots$?

there will always be some properties for which no mechanical short cuts are possible. Hence no mechanical rule can always answer the question correctly. Furthermore, we observe that a formal (i.e. mechanically checkable) proof of a negative answer would automatically provide a short cut. This forms the basis of Kleene's proof (mentioned by Kreisel in § 6 below) of Gödel's incompleteness theorem: *in any formal system that is adequate for number-theory there will be true statements* (of the form 'no number has *P*') *that cannot be proved.*

This leads naturally to the question whether the human mind can transcend the limitations of computers. It is not easy to discern a line of argument that could conceivably settle this question. Of course, the flexibility of the mind when faced with a problem is dazzling; it can select an appropriate known rule, create a new one, or work unsystematically. But computers also can be given these abilities. The brain has been evolving for millions of years, human knowledge for thousands of years. The progress that digital computers have made in less than forty years gives little warrant for supposing that they cannot match the creative power of the human intellect.† Nor is there any reason to suppose that one computer could not design and create another.

But the existence of mechanically unsolvable problems does at least enable one to *picture* the transcendence of minds over machines. One must imagine that as mathematics develops mathematicians will produce unsystematically solutions to an ever-increasing number of cases of an unsolvable problem. But this *is* only a picture; it could never be an established fact. For the production of solutions must be unsystematic, since the problem was supposed unsolvable. So there can never be a guarantee that the production can always be continued. Finally, it should be remarked that a transcendence of the kind considered need not be unphysical. For although the consequences of an acceptable physical theory must be by and large computable, there may be certain special initial conditions whose consequences cannot be computed.

(b) Foundations of Mathematics: 1900–1950

Foundations provide a theory of mathematical practice. Thus, in the first place, they describe and analyse our mathematical experience, that is the body of results and methods as they present themselves to the

† Who would have predicted thirty-five years ago that a machine would play draughts at national championship standard and have a style of play that other players could recognize? Yet this is now the case.

working mathematician; as any other theory, foundational results lead to an extension of mathematical practice. *Logical* foundations, which are treated here, provide a logical analysis, that is, they are principally concerned with questions of meaning and validity.

The period considered is sufficiently distant to deserve the attention of the general reader: the fads and fashions among the specialists working at that time can be seen in perspective. Also, since actual *knowledge* of an established subject is usually a pretty good measure of depth of genuine interest, the reader of this survey may be expected to have a little previous knowledge. Actually, used imaginatively, quite *elementary* parts of school mathematics illustrate several of the abstract ideas used in the text below.

In connection with *sets* and their properties it is best to think not primarily of geometric figures, but of combinations and permutations, and the analysis needed to find arithmetic laws about such combinatorial operations on sets of objects (whose nature is quite immaterial); for instance that there are 2^n ways of picking, or not picking, members of a collection, say C, of n objects. Each way determines a subcollection of C, consisting of the objects that have been picked, and the 2^n subcollections (counting also the empty 'set') form what is called the *power set* $\mathfrak{P}$(C) of C.

In connection with *reductions* and with *definitions* in simpler terms it is best to put oneself in the place of someone who has formed geometric concepts such as: circle, square, or, to be a little fancier, convexity, and then *discovers* an analysis of these concepts in terms of a very *few* geometric notions such as: equidistant and collinear; for instance, he discovers that the figures that he has come to recognize as circles are the *loci* of points that are at the same distance from some given point (the centre of the circle).

In connection with *non-constructive* definitions one may think of such simple matters as forming the sum of two decimal expansions, given by rules which allow one to calculate each decimal place quite mechanically, say $0{\cdot}a_1\ a_2\ \ldots$, $0{\cdot}b_1\ b_2\ \ldots$. They have a sum, say $\mathrm{c}_0{\cdot}\ c_1\ c_2\ \ldots$: can we find c_0? For any given n, however large, we may not know if $c_0 = 0$ or $c_0 = 1$, e.g. if $0{\cdot}\mathrm{a}_1 \ldots \mathrm{a}_\mathrm{n}$ and $0{\cdot}b_1 \ldots b_\mathrm{n}$ add up to $0{\cdot}9 \ldots 9$; 'tiny' changes in the given rules may make c_0 jump from 0 to 1 and conversely. Constructivity is concerned with processes and what we know about them, not only the final result. Only some of the more delicate points (see the end of the text) can, demonstrably, not be *concretely* illustrated by elementary mathematics.

One final word of introduction: despite its decisive role, the concept of computability will be somewhat neglected in the survey below; the reader is referred to Professor Gandy's article. The same applies to formal derivability, which is defined in terms of (mechanical) computability.

By now it will be a relief for the reader to plunge *in medias res.*

1. *Legacy of the nineteenth century.* Cantor studied the new and fundamental mathematical concept of *abstract set*; fundamental because the mathematical notions studied previously can be defined set-theoretically (natural numbers by the work of Peano and Dedekind; the continuum by that of Dedekind). The set-theoretic development of mathematics brought into prominence *non-constructive* methods of reasoning corresponding to operations on infinite sets;† earlier reasoning had a more computational character (Kronecker stressed the difference). A link between the two kinds of reasoning was restored by the *formalization* of predicate logic (Frege): all intuitively valid statements of so-called elementary logic known at the time are generated by certain purely mechanical rules, and these rules produce only valid statements. Using the symbolism of predicate logic Hilbert formulated the notion of *geometric* (Euclidean) *proof* and gave independence results (the parallel axiom) in a *formally precise* manner.

Philosophically, for the first time the possibility of a systematic development of all mathematics seemed feasible thanks to the expressive power of the simple vocabulary of set theory and elementary logic. Also for the first time, positivistic epistemology gained a foothold in mathematics thanks to formalization, which 'replaced' the abstract notion of logical consequence by the concretely realizable relation of formal derivability

2. *The first decade of the twentieth century: big ideas.* As early critics of Cantor had foreseen, the notion of *set* needed analysis (paradoxes). After some intermediate stages one arrived at *one* analysis, considering as sets the objects obtained by iterating the following construction: given such an object A, for example the empty set, one forms the collection $\mathfrak{P}(A)$ of all *subsets* of A and the union of A and $\mathfrak{P}(A)$ (full cumulative hierarchy of sets); Zermelo gave simple properties of the structure obtained by iterating up to so-called limit ordinals (Zermelo's axioms).‡

† If we were to use the logical law: either (i) all natural numbers have the property *P* or (ii) some number does not have the property *P*, as a *rule* [do *X* if (i) and do *Y* if (ii)], an *infinite* search would be involved; cf. R. O. Gandy's contribution.

‡ To spell it out, one starts with A_0, takes for B_0 the union of $A_0, A_1, \ldots$, where $A_{n+1} = A_n \cup \mathfrak{P}(A_n)$; for C_0 the union of $B_0, B_1, \ldots$, where $B_{n+1} = B_n \cup \mathfrak{P}(B_n)$; and so on; or, using the notation of *ordinals* where ω is the first infinite ordinal, $B_n = A_{\omega+n}$, $C_0 = A_{\omega+\omega}$

Russell and Poincaré gave a more *refined* analysis: instead of considering *all* subsets of A only those *definable* from A in the vocabulary of elementary logic are used (ramified hierarchy).

Whitehead and Russell (*Principia Mathematica*) present in some detail the *reduction* of mathematics to set theory, giving not only definitions associating to an ordinary mathematical statement A, say in arithmetic, its set theoretic *translation* A_S, but listing the properties of sets (axioms) that imply A_S logically whenever A had been proved mathematically.

Hilbert proposed a programme for *justifying* the positivistic position within mathematics, roughly speaking by showing that truth and formal derivability coincide. (For the case of elementary arithmetic this reduces to his consistency problem: in terms of the preceding paragraph, A is true if and only if A_S is formally derivable.) This programme was to kill two birds with one stone by exhibiting the adequacy of formal systems and by showing, at the same time, the validity of elementary conclusions obtained by non-constructive reasoning: for the latter purpose, the consistency proof had to be given by elementary methods, which Hilbert called *finitist* (since he regarded 'finiteness' as essential to their elementary character).

Brouwer, unlike Hilbert, proposed to *ignore* non-constructive reasoning altogether because he considered it as based on an error (uncritical extension of principles valid for concrete finite collections). But, again unlike Hilbert, he considered not only finitist operations but also constructions on more abstract objects, roughly speaking on constructions that, in turn, refer only to our own mathematical constructions. For example, he considered constructions on intuitive *proofs* and spoke of intuitionistic mathematics; he stressed particularly that the law of the excluded middle is not valid (when interpreted in terms of constructions on proofs).†

The development of each of the leading ideas mentioned in Section 1 will be sketched below. Speaking chronologically, the decade after the

† Intuitionistic rules are more subtle than the *mechanical* ones discussed by Gandy above, as can be seen by comparing formal derivations and convincing arguments (which the derivations are intended to represent). We use the formal rule: from A and $A \rightarrow B$ derive B, *because* we have seen that a valid proof of A and a valid proof of $A \rightarrow B$ yields a valid proof of B; just being a 'formal rule' does not make it valid (as illustrated by inconsistent rules). The formal rule tells us to compound two formal objects (derivations) to get a third; the (intuitionistic) rule, which makes the formal rule valid, tells us how to combine two valid proofs to get a third. The latter would be useless to a machine which operates only on the concrete representation of proofs.

First World War consisted largely of consolidation and technical development, while the thirties pin-pointed some of the real problems involved in the big projects. (The difficulties discovered were not surprising, because, from a naïve point of view, each project *was* daring when first proposed.) Perhaps because of these difficulties, the last decade (1940–9) was most notable for the *application* of earlier foundational research, both to various branches of mathematics and to computers; their relevance to the present report is merely that they provided a *test* of our foundational understanding.†

3. *Cumulative hierarchy of sets*. Perhaps because initial progress in set theory had been very rapid, further progress was slow. Mahlo, Fränkel, and Zermelo discovered some properties, nowadays called *axioms of infinity*, of the hierarchy obtained by iterating the basic construction in (2) sufficiently often (i.e. to large ordinals). Gödel pointed out that such properties allow one to derive, formally, new number-theoretic propositions. Tarski gave an explicit set-theoretic *reduction* of the notion of logical consequence itself. On the negative side he noted an *inadequacy* of the language of set theory (indefinability of the truth predicate) because certain collections of natural numbers cannot be defined in this language when it is interpreted as intended. Gödel, in the late forties, pointed out that, as far as statements in the language of set theory are concerned, the effect of this inadequacy is offset by use of suitable axioms of infinity. At the same time he stressed, eloquently but with little immediate effect, the need for discovering new axioms, in other words for continuing the process that led to the existing axioms of set theory, and mentioned Cantor's continuum problem as a possible example whose solution may require such axioms. (There was some experimentation, mainly in Poland, with properties 'analogous' to those of ω, the first infinite ordinal, but their character as axioms of infinity was realized only recently.)

4. *Intuitionistic mathematics* took rapid strides in the twenties, when Brouwer gave a uniform constructive theory of ordinals (transfinite

† Actually, these early applications did not provide a *convincing* test because they were relatively simple; but they led to recondite applications in the second half of the twentieth century. Applications of the general theory of axiomatic systems (structures defined in elementary predicate logic) begun in the thirties by Birkhoff and, above all, Malcev, were continued by Henkin, Malcev, and Tarski and above all by Abraham Robinson. Applications of recursion theory to word problems in algebra were made by Post and Turing, and applications of the methods of proof theory to extract explicit bounds from *prima facie* non-constructive proofs in number theory by G. Kreisel.

iteration of constructive operations), of free choice sequences and operations (functionals) defined on them, and of (intuitive) proofs consisting of the transfinite iteration of basic steps. Heyting, in the thirties, was the first to attempt a coherent analysis of *constructive logical* operations in terms of the basic notions of proof and construction. His formulation made the problematic and abstract character of these operations evident. On the formal side, Kolmogorov, Gentzen, and Gödel established simple relations between derivability in the then current non-constructive and intuitionistic formal systems for elementary logic and number theory. In the forties Brouwer published some fragmentary ideas extending the range of intuitionistic mathematics by analysing explicitly what the (idealized) thinking subject or mathematician can know about his own reasoning.

5. *Definability* (in connection both with the abstract hierarchy of sets and with the abstract constructions treated in intuitionistic mathematics). Recall first the Russell–Poincaré hierarchy of Section1. Gödel considered a kind of hybrid, using the Russell–Poincaré definition principle, but defining its transfinite iteration in terms of Zermelo's notions (so-called set theoretic von Neumann ordinals). He showed that the objects of his hierarchy (which he called *constructible*) satisfy not only the axioms originally discovered for the full cumulative hierarchy and hence all the conclusions drawn from them, but also that several open problems such as the continuum hypothesis can be settled if one confines oneself to constructible sets. Generally speaking, as far as propositions in the usual language of set theory are concerned, we must either look for new properties of the cumulative hierarchy to distinguish formally between it and that of Russell–Poincaré, or else formulate the latter more sharply.

In intuitionistic mathematics the best known kind of *definable* operations (at least among those applied to natural numbers) are the *recursive* functions; and, similarly, proofs described by derivations in suitable (recursive) *formal systems* are the definable 'part' of the abstract proofs of Brouwer's intuitionistic mathematics. Church and Kleene showed how to develop Brouwer's definitions of ordinals when one restricts his functions to recursive ones, and Kleene showed that the formal principles formulated in (so-called Heyting's first-order) arithmetic are satisfied if the constructions, more or less implicit in the intuitionistic logical operations, are taken to be recursive.

6. *Hilbert's programme* produced the most conclusive results during the period considered. On the positive side, substantial portions of current non-constructive number theory were reduced, in Hilbert's sense, to

evidently finitist principles (in most polished form by Gentzen). In the same direction Gödel's *completeness* theorem established mathematically the *general* idea behind Hilbert's programme, in the area of elementary logical reasoning (about truth functions and quantifiers; Frege's original discovery of the rules of predicate logic, fifty years earlier, was only empirical, justified by a case study). In contrast, the general idea was suspect when Loewenheim realized that each of the principal axiomatic systems considered by Hilbert is satisfied by several quite distinct structures or models. However, since Hilbert's general idea considered assertions (formulae) and not structures it was rigorously refuted only by Gödel's incompleteness theorems; specifically, refuted as far as reasoning of a genuinely mathematical kind, for instance number theory, is concerned. Gödel himself considered formal derivations in a version of *Principia Mathematica*, but afterwards this was extended by Kleene to *arbitrary* formal systems (in the sense that for each formal system containing a minimum of arithmetic there is a recursive function with the property: none of the recursion equations that define it can be formally proved to define a function at all). The generalization depended on Turing's surprisingly simple and convincing analysis of what a formal system or mechanical procedure is. (For generalizing Gödel's 'negative' result the analysis was essential, but not directly for Hilbert's original programme.)

A common reaction to the incompleteness theorem was to regard Hilbert's programme as having failed without giving up the idea behind his programme! (neo-formalist foundations). Indeed some went so far as to give up the idea of *any* logical foundation or analysis of the validity of principles and began to speak of 'foundations', for example 'foundations of ring theory', simply in the sense of a well-organized exposition.† But a few years after the incompleteness theorem Gentzen discovered a reduction of arithmetic to very *elementary* uses of so-called ϵ_0-induction (not using the abstract operations implicit in classical or intuitionistic logic). Though he did not analyse exactly in what this elementary character consisted, inspection of his work makes it clear that *some* substantial reduction had been achieved; Hilbert's programme was viable provided one replaces his finitist proofs by a suitable larger, but significant,

† Presumably because they thought that this was the only reasonable enterprise that got anywhere near the traditional idea of foundations. Forgetting that Hilbert's programme was a quite specific and not at all naïvely plausible foundational scheme. Evidently this kind of reaction was not likely to advance the subject directly; worse still, it denied whatever progress had been made already.

subclass of constructive proofs. So, contrary to the popular opinion mentioned, Gödel's incompleteness theorems refute the general idea behind Hilbert's programme, but not the possibility of modifying the programme fruitfully.

After Gentzen, proof theory, as Hilbert called work on his programme, consisted of more isolated results, not surprisingly concentrating on methods and notions closer to intuitionistic than to Hilbert's or Kronecker's finitistic mathematics. These were infinite proof figures (Novikov and Lorenzen); constructive functions of finite type, i.e. one has constructive number theoretic functions given by laws, operations on such laws, operations on such operations, etc. (Gödel, but not published till 1958); and constructive functionals on freely chosen sequences (Kreisel). In the last two studies, as a further departure from Hilbert's original programme, one considers not only formal translations of quite elementary numerical statements but assigns a *constructive meaning* (interpretation) to the whole non-constructive formalism.

7. *Principal foundational problems* arising from the work described above (as judged in 1968). For development of Cantor's notion of set or, more exactly, of the cumulative hierarchy, one was waiting in 1950 for the discovery of new properties, preferably those evident for 'long' segments of the hierarchy (axioms of infinity), and, specifically, for old problems of number theory which can be decided by the use of such properties. In others words, instead of looking at the integers as part of the complex or *p*-adic numbers (as is familiar from analytic number theory and its modern modifications) and deriving recondite arithmetic laws from simple laws satisfied by the larger structure, one embeds the integers in the 'huge' full cumulative hierarchy. Gödel's results made this step promising, but not more than that.

The state of intuitionistic mathematics in 1950, with its diverse notions and *ad hoc* methods, was comparable to that of classical mathematics before systematic set theoretic foundations had been given: Heyting's formal systems for elementary intuitionistic logic had the empirical status of Frege's predicate calculus; no analogue to Gödel's completeness theorem was in sight in 1950 (for completeness with respect to the Brouwer–Heyting interpretation of logical operations). And continuing the parallel with set-theoretic foundation, one had still to find, and analyse precisely, important *parts* of intuitionistic mathematics, that is important *kinds* of constructions and evidence (proof)—areas as significant for constructive foundations as, for instance, the (hereditarily) finite sets are for set-theoretic foundations. Hilbert's notion of finitist proof and Bouwer's transfinite constructions by use of (his) ordinals were clear candidates for such areas.†

To make definability results described in Section 5 philosophically more meaningful, the use of the abstract notion of set-theoretic ordinal, respectively of constructive function, must be eliminated. The principal missing element appears to be some *elementary* notion of transfinite iteration (perhaps related to the abstract notion of iteration as polynomials are to abstract functions); Gentzen's use of ϵ_0-induction mentioned in Section 6, when properly analysed, is probably a typical illustration.

Further development of proof theory must face two problems (which Hilbert had hoped to avoid). First, since by Gödel's incompleteness theorem *all* formal systems are incomplete, i.e. inadequate with respect to truth, a subtler criterion than completeness is needed for a rational choice of formal systems worth studying. Second, since reduction to finitist methods (whatever they may turn out to be) is not likely to be generally possible, a meaningful choice of methods is needed. Presumably this will be provided by a solution of the last problem above on intuitionistic foundations.

8. *Some final comments.* There are some outstanding foundational problems to which research in the first half of the present century (or, for that matter, since) has contributed very little. At one extreme we are still waiting for an analysis of the idea of *arbitrary* iteration (not only to a specific ordinal) in connection with the cumulative hierarchy. This idea, which certainly cannot be analysed in set theoretic terms, may well use a very abstract notion of function that can be applied to itself, a notion dimly involved in the paradoxes. So to speak at the other extreme, we have the idea, so central for *actual* understanding, of what is graspable; in particular, to be graspable, proofs must not only be valid but intelligible, and configurations must not only be finite but it must be possible to visualize them.

At the present time we do not have the notions to formulate these ideas precisely. But, and, to my mind, this is one of the most striking aspects of the work described above, time and again traditional epistemological conceptions and ideas have turned out to have not only a precise but an essentially unambiguous formulation, at least after a few preliminary distinctions had been made. (Perhaps not unlike the passage from the Greek conception of atomic structure to modern atomic theory, or from

† Significant progress has been made in the last twenty-five years with the problems mentioned in the last two paragraphs.

the ideas of fire, earth, water, and air to heat/light, solid, liquid, and gas respectively.)

Going further, even some of the more extravagant doctrinaire positions, which set implausible *limits* to reasoning or acceptable evidence, have proved to be fruitful, provided only that one regards them not negatively as restrictions, but positively as drawing attention to particularly significant and internally coherent *kinds* of reasoning or evidence.

Lastly, as for instance the success of Hilbert's programme in elementary mathematics and its failure in more advanced branches shows in connection with positivism, some traditional philosophical distinctions can be illustrated *concretely* only by going outside the mathematical knowledge of the 'educated common man' (at the present time or even of the people who first made the distinctions). Here it may be remarked that even if the 'new maths' does not turn out to be as useful, practically or pedagogically, as is sometimes claimed, it certainly provides good illustrations for certain traditional *philosophical* ideas that are not illustrated by material found in the old *mathematical* school curriculum.

General References

1. J. van Heijenoort, *From Frege to Gödel*, Cambridge, Mass., 1967, a most scholarly anthology of basic works in logic up to the early thirties.
2. E. Zermelo (1930) *Fundamenta Mathematicae* **16,** 29–47. The article *Über Grenzzahlen und Mengenbereiche* gives an unsurpassingly clear foundation of set theory not included in 1.
3. Martin Davis, *The Undecidable*, New York 1965, an anthology covering most of the principal work in the thirties and forties, but excluding Gödel's work on the Russell-Poincaré (*constructible* or *ramified*) hierarchy.
4. P. Benacerraf and H. Putnam, *Philosophy of Mathematics*, Englewood Cliffs, New Jersey, 1964. An uneven collection but includes Gödel's own views on the implications of his work on the constructible hierarchy.

2 Relativity and Cosmology

1. Introduction

THE sudden flowering of the theories of special and general relativity in this century came about through the union of two hitherto unrelated concepts, that of the relative character of motion and that of the relationship between mechanics and optics. These two together led to the special theory of relativity, and when we use this theory to discuss gravitation we are led in turn to the general theory of relativity. This general theory has made it possible to give a discussion of the behaviour of the universe on the largest possible scale, that is, of cosmology, in a much more satisfactory way than was possible before 1900.

Galileo,[1] in 1632, had imagined the thought-experiment of observing flying insects, fish in bowls of water, and so on, in a uniformly moving ship. The equivalence of rest and uniform motion shown by this was taken over by Newton[2], although with the *caveat* that an absolute space and time exist but are not easily discoverable by mechanical experiments.

The earliest discussion of the relation between mechanics and optics seems to have been that of Euler[3] (1750). He noticed that the investigation of aberration (the change in apparent position of the stars due to the Earth's motion, measured by Bradley[4] in 1727) gave different answers according to whether we consider a ballistic or a wave theory of light. For the ballistic theory the 'moving Earth, fixed star' system is equivalent the the 'fixed Earth, moving star' system shown in Fig. 2.1, giving an angle of aberration α where

$$\tan\alpha = \frac{v\sin\phi}{c - v\cos\phi}.$$

For the wave theory the two systems are not equivalent (the medium must be taken into account), and the corresponding formula is different.

This investigation of Euler's was by far the earliest of a number of later discussions, and it will be sufficient to consider here an experiment conducted by Fizeau[5] to test a theory of Fresnel.[6] This experiment is not one that can be carried out with very high accuracy (for such experiments it is better to take that of Michelson and Morley[7]), but it has the advantage of great simplicity in its theoretical description and in its explanation.

Light is passed through a long pipe through which water flows. When the water is at rest, the speed is c/n where n is the refractive index. When the water flows with speed v, it is

$$\frac{c}{n}+v\left(1-\frac{1}{n^2}\right),$$

a result derived by Fresnel by an acoustic analogy.

Little further progress can be made without extending the theoretical foundations. The late nineteenth century saw the development to a very high degree of the theory of electromagnetism, culminating in Maxwell's theory of the electromagnetic field.[8] In passing it may be noticed that Riemann[9] had observed the experimental agreement between the speed of light in free space and the quantity $1/\sqrt{(\mu\kappa)}$ which arises in electromagnetic theory in transforming from electrostatic to electromagnetic units. He had also observed that when one takes account of this speed it is possible to generalize the equation satisfied by the electrostatic potential to the form of a wave equation, describing wave transmission with this velocity. But he had no theoretical basis for this equation. With Maxwell's theory, such a wave equation appears automatically, and this constitutes its greatest achievement.

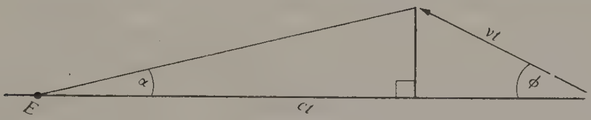


FIG. 2.1

It will be necessary to describe Maxwell's theory briefly in order to understand the later developments. Maxwell deals with four quantities (directed quantities, i.e. vectors) for describing the state of the electromagnetic field at any point. These are the electric and magnetic field-strengths **E** and **H**, and the electric and magnetic inductions **D** and **B**. In free space we have the equations

$$\mathbf{D}=\kappa\mathbf{E},$$
$$\mathbf{B}=\mu\mathbf{H}.$$

Maxwell's first relation is the usual law of magnetic induction that the electromotive force in a closed circuit is proportional to the rate of change of the flux of **B** from the circuit (with a negative sign (Lenz's law))[10] Maxwell then considers the corresponding situation with the magnetomotive force; the situation here, however, is much more complicated. In the first place the magnetic induction is a special kind of field which is conserved; that is to say, it has the same form as the velocity field in an incompressible fluid, which is such that the amount of fluid inside a volume can increase only by fluid flow across the surface. Such a field is one for which it makes sense to consider its flux *through* a circuit. Maxwell would have liked to have had a corresponding law for magnetomotive force in terms of the rate of change of the flux of electric induction, but the electric induction field has sources, to wit, the charges in the field. Accordingly Maxwell had to modify the equation by adding a term which is such that the total quantity whose flux is being taken *is* a conserved field (since it is only for a conserved field that this concept of flux has a meaning). This involves adding to the rate of change of electric induction a term involving the currents flowing (it may be noted that Maxwell's original approach was rather the opposite of the one given here; the term involving currents was experimentally verified, and he then added the rate of change of electric induction in order to produce a conserved field). It is a purely mathematical exercise to show that the fields which satisfy these equations satisfy also the wave equation with the speed noticed by Riemann; so amongst the possible fields are some propagating with wave motion equal to that of light. This then is the electromagnetic theory of light which provided the foundation for the later development of optics.

This theory predicts a unique speed for light, independent of any motion of the light source, or of the observer. Such a unique speed seems to select, from amongst the frames of reference suitable for Newtonian mechanics, a unique one—Newton's absolute space. But by the nineteenth century such a situation was no longer tenable; the so-called inertial frames (those in which Newton's laws have particularly simple forms) were known to be completely equivalent mechanically speaking.[11]

The solution to this problem turns out to be a slight alteration in the definition of transformation between inertial frames. The way in which this transformation was carried out in Newtonian mechanics was

$$x' = x - vt, \quad t' = t.$$

These two equations state two independent facts about the system. The first tells us that we are using a coordinate system whose origin is moving uniformly relative to the old one (Fig. 2.2). The second states that the same time-variable is used in each coordinate system; the importance of this is that if two events are simultaneous in one coordinate system they are simultaneous in all other inertial frames. Put in this form the second equation represents a physical assumption, although one whose importance was not at first realized. It assumes that we have some means of determining which events are simultaneous. In the solution of the problem of transforming Maxwell's equation from one inertial frame to another it turns out that this principle of simultaneity has to be given up.

The detailed solution of this problem for Maxwell's equations was given by Poincaré[12] and Lorentz[12] in 1904. However, neither of these authors would have expressed his theory in the way we have just used.

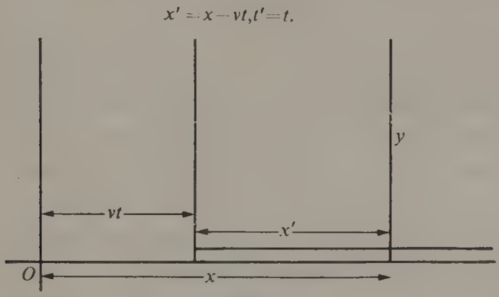


FIG. 2.2

Both of them gave sets of equations for transforming the Maxwell field quantities and the supposed time-coordinates in such a way that Maxwell's equations were unchanged under the transformation. In fact the solution they found to the problem was the so-called *Lorentz transformation.*

$$x' = \beta(x - vt), \quad t' = \beta\left(t - \frac{vx}{c^2}\right), \quad \beta = \frac{1}{\sqrt{(1 - v^2/c^2)}}.$$

It is clear from these equations that, so long as the relative velocity of the reference frames is small compared with that of light, the transformations will be indistinguishable from those of Newton, and accordingly the new mechanics, which is unchanged under the new transformations, will differ only by very small terms from the mechanics of Newton. At the same time it must be emphasized that neither Poincaré nor Lorentz had in mind at the time the essential physical assumption here that distant simultaneity had to be defined. Indeed Lorentz did not believe that the new time-variable introduced by these equations was the time measured in the moving reference frame. He wrote 'But I never thought that this had anything to do with real time . . . there existed for me only one true time. I considered my time-transformation only as a heuristic working hypothesis'. We shall take up the problem of distant simultaneity and its solution by Einstein in the next section.

The other ingredient of general relativity was the theory of gravitation. This theory had been set forth with great clarity and completeness by Newton[2] in the *Principia* and was based on the experimental fact that the force of attraction between any two particles is proportional to the product of their masses and inversely proportional to the square of the distance between them. The fact that the attraction is proportional to the product of the masses had been known for very much longer in a slightly different form, for since, according to Newton's mechanics, the force is the product of the mass and the acceleration, it follows that two different masses in the gravitational field of a third move with the same acceleration when they are in the same position, although their masses are different. This fact was known to Galileo. (For example, in 1638[13] he said 'Aristotle said that an iron ball of 100 lb falling from a height of 100 cubits reaches the ground before a 1-lb ball has fallen a single cubit. I say that they arrive at the same time'). The fact that the law of force is as the inverse square of the distance is Newton's unique contribution, and and one that came to be much emphasized between the acceptance of Newton's theory and the end of the nineteenth century. With the help of this result of Newton's, astronomy prospered, and the orbits of the planets round the Sun were determined with very great accuracy.

Towards the end of the nineteenth century, however, some small discrepancies were found between the predictions of Newton's theory and the observed motion of the planets. The most significant of these concerned the advance of perihelion of the planet Mercury. Newton's law of gravitation predicts that a single particle moving round a body like the Sun will do so in an orbit that is an ellipse with the Sun at one focus. We do not have to deal, however, with a single body but the whole collection of planets moving round the Sun. Each of these planets attracts all the others to itself with very much smaller forces than the Sun's force, and the effect of this turns out to be that the elliptical orbit due to the Sun slowly rotates. The observed advance in perihelion is 5600″ (seconds of arc) per century. The calculated advance on the basis of the perturbation of the other planets is 5557″. The discrepancy of 43″ per century appears to be inexplicable on a Newtonian basis. The advantage of a modified mechanics as a result of the theory of Poincaré and Lorentz naturally suggests that when gravitation can be included this existing advance of perihelion will also be provided by the theory. The problem of including gravitation proved, however, much more difficult than was originally supposed, and the solution of it is entirely Einstein's.

2. Einstein's contribution

Einstein brought to the problem tackled by Poincaré and Lorentz a completely fresh approach. He considered[14] the operational procedure for assigning time to distant events, noticed that it *must* involve some kind of signalling, and accordingly used the unusual character of the speed of light as an argument for defining the time by means of light-signals. Since the time to be assigned to a distant event is a *convention*, he is at liberty to assign to it the *average* of the times of emission of a signal and reception of the reflected signal.

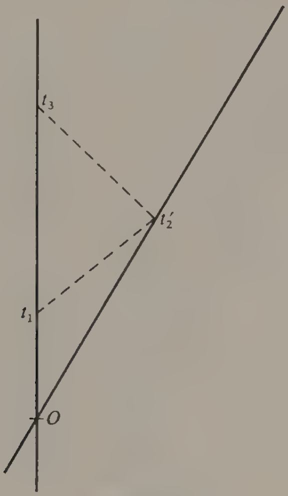


FIG. 2.3

Let us then see how Einstein's rule may be made consistent with the remark (made above) that all inertial observers give equivalent descriptions of events. Suppose that two inertial observers are considered, and one O' is moving uniformly with speed v relative to the other O. We can draw a diagram (Fig. 2.3) to represent their relative motion, in which the vertical axis represents the time as measured by O and the horizontal axis the distance travelled. It is convenient in what follows to adopt the speed of light as the unit of speed, so that the space and time axes have a common scale and the path of a light ray will be a line at 45° to the time axis. Fig. 2.3 shows the signal we have been discussing, by means of which O will determine the whereabouts of O' at a certain time, the event in the above discussion being the instant of reflection of the signal. The time t_2' is the time assigned to this event by O', an assignment that he could make directly, not by using Einstein's rule, since the event happens to himself. It is clear that t_2' will be proportional to t_1; and so, inserting a constant of proportionality, k, which will depend on the relative motion,

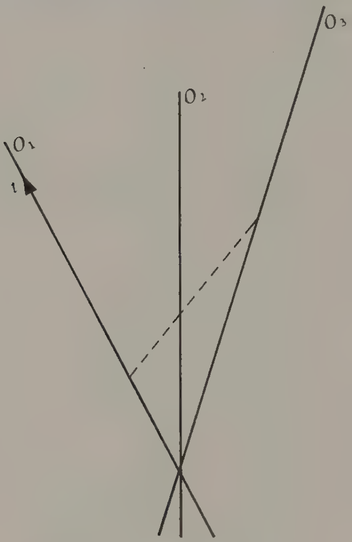


FIG. 2.4

we have $t_2' = kt_1$. But from the symmetry of the two equivalent observers we have equally $t_3 = kt_2'$, from which it follows that

$$t_3 = k^2 t_1, \quad t_2 = \tfrac{1}{2}(k^2+1)\,t_1 = \tfrac{1}{2}\left(k + \frac{1}{k}\right) t_2',$$

$$x = \tfrac{1}{2}(k^2-1)\,t_1, \quad v = \frac{x}{t} - \frac{k^2-1}{k^2+1}.$$

These equations contain all the essential features of special relativity.[15]. The constant k is determined in terms of the speed of separation and the different times assigned to the reflection event are given in terms of it.

These formulae enable us to derive a result that explains the experiment of Fizeau[5] described in the last section. Consider the case of three observers (Fig. 2.4), one of whom, O_1, determines the coordinate system and so remains at rest in it, and the other two of which move uniformly relative to O_1, all three being coincident at $t=0$. If the ratio of times of a light signal at O_1 and O_2 is k_{12}, and that for a signal from O_2 to O_3 is k_{23}, it is clear by considering a signal sent from O_1 to O_3 passing through O_2 that

$$k_{13}=k_{12}k_{23}.$$

Expressed in terms of relative velocities by means of the expression of k in terms of speed this gives at once

$$v_{13}=\frac{v_{12}+v_{23}}{1+v_{12}v_{23}},$$

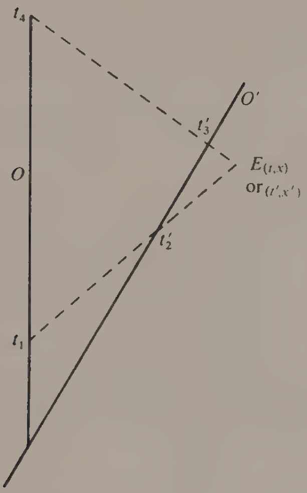


FIG. 2.5

in agreement with the Newtonian expression for relative velocity, so long as the speeds concerned are small compared with that of light. In the Fizeau experiment the equations become

$$v_{13}=\frac{(1/n)+v}{1+(v/n)}\simeq\left(\frac{1}{n}+v\right)\left(1-\frac{v}{n}\right)$$
$$\simeq\frac{1}{n}+v\left(1-\frac{1}{n^2}\right),$$

in complete accordance with the observed values. It is noteworthy also that the result of combining any speed with that of light according to the new relative velocity formula is to give the speed of light again, so that this speed does indeed have the special characteristic that we would expect in the theory. Moreover, the equations for the transformation from one observer to another when written in terms of space and time coordinates do have the form of the Lorentz transformation mentioned in section 1, which had already been shown by Poincaré and Lorentz to be the form of transformation that left Maxwell's equations unchanged. Accordingly, Maxwell's equations and Einstein's convention for assigning time to distant events are completely consistent.

It is necessary, in order to understand later developments, to carry out this rewriting of Einstein's theory in terms of orthodox time and space coordinates;[16] when this is done the transformation from one observer to the next can be derived as follows (Fig. 2.5).

The equations

$$t = \tfrac{1}{2}(t_4 + t_1), \quad t' = \tfrac{1}{2}(t_3' + t_2'),$$

$$x = \tfrac{1}{2}(t_4 - t_1), \quad x' = \tfrac{1}{2}(t_3' - t_2'),$$

may be written

$$t_1 = t - x, \quad t_2' = t' - x', \quad t_3' = t' + x', \quad t_4 = t + x.$$

Hence

$$t' - x' = k(t - x),$$

$$t' + x' = \frac{1}{k}(t + x),$$

giving

$$t' = \frac{1}{2}\left(k + \frac{1}{k}\right)(t - vx) = \beta(t - vx)$$

$$x' = \beta(x - vt),$$

where $\beta = \frac{1}{2}\left(k + \frac{1}{k}\right) = (1 - v^2)^{-\frac{1}{2}}$, results, as mentioned above, known to Poincaré and Lorentz.

This treatment has, it is true, been wholly one-dimensional. It is next assumed that the effect of a translation of the coordinate axes is only in the direction of the translation, so that for three dimensions the transformation takes the form, say,

$$x' = \beta(x - vt), \quad y' = y, \quad z' = z$$

$$t' = \beta(t - vx),$$

where the x-axis has been drawn in the direction of separation. The resultant of two such transformations, say along the x and y axes, is not another transformation of the same kind. One could show this by carrying out the transformations explicitly; it is a little easier to do it by beginning with a set of axes $Ox_1y_1z_1$ and defining the terms as follows.

$Ox_2y_2z_2$ is derived from $Ox_1y_1z_1$ by translation, velocity u, along Ox_1,

$Ox_3y_3z_3$ is derived from $Ox_2y_2z_2$ by translation, velocity v, along Ox_2,

$Ox_4y_4z_4$ is derived from $Ox_3y_3z_3$ by translation, velocity $-u$, along Ox_3,

$Ox_5y_5z_5$ is derived from $Ox_4y_4z_4$ by translation, velocity $-v$, along Ox_4.

The resultant of these transformations is not the identity transformation; for even if we retain only terms of order v^2 we have, approximately, amongst other formulae,

$$x_2 = (1+\tfrac{1}{2}u^2)\,x_1 - ut_1, \quad t_2 = (1+\tfrac{1}{2}u^2)\,t_1 - ux_1,$$

$$x_3 = x_2, \quad t_3 = \{1+\tfrac{1}{2}(v^2+u^2)\}\,t_1 - ux_1 - vy_1,$$

$$x_4 = (1+\tfrac{1}{2}u^2)\,x_2 + ut_3,$$

$$= x_1 - uvy_1,$$

with similar formulae for the other coordinates. Thus x_4 is derived from x_1, not by the identity transformation, but by a rotation through a small angle of about uv radians. (This is the so-called *Thomas precession*,[17] at one time put forward as an explanation of electron spin.) The lesson to be drawn is that it is confusing to consider such transformations alone; they should be taken in conjunction with the rotations of space.

The most useful mathematical trick for doing this was largely created by H. Minkowski,[18] who suggested the use of four-dimensional geometry for the purpose. The transformations of coordinates have the property that the quantity $t^2-x^2-y^2-z^2$ is left unchanged by them. Since $x^2+y^2+z^2$ is distance from the origin, it, and t separately, are left unchanged by space rotations, and so rotations also have the property of leaving $t^2-x^2-y^2-z^2$ unchanged. It is convenient to introduce the notations $t=x^0$, $x=x^1$, $y=x^2$, $z=x^3$, so that the coordinate transformations can be written

$$x^a \to x^{a'} = f^{a'}(x^a) \quad (a, a' = 0, 1, 2, 3)$$

$$= \sum_a l_a^{a'}\, x^a,$$

where $l_a^{a'}$ are a set of constants. Formulae of this sort may be much shortened by a systematic use of the summation convention, due to

Einstein,[19] that an occurrence of a repeated literal suffix, as *a* here, shall be taken as an instruction to sum over the values 0, . . . , 3 of the suffix. The transformation may now be written

$$x^{a'} = l_a^{a'}\, x^a$$

and the quantity left unchanged may be denoted by $\eta_{ab}\, x^a x^b$, where $\eta_{00} = 1$, $\eta_{11} = \eta_{22} = \eta_{33} = -1$, $\eta_{ab} = 0$ otherwise.

3. The principle of relativity†

The fact that this transformation is different from the Newtonian one, and the necessity for finding the quantities that are not changed by the transformation, led to a more careful analysis in the form of the principle of relativity. This achieves its most explicit recognition in the work of Hermann Weyl.[20] We have to do with the assigning of numbers to certain physical systems by means of an experiment, and the experiment is specified in a certain coordinate system. Our requirement is to find those numbers that are genuine properties of the physical system rather than of the coordinate system alone. The technique for doing this is to consider transformations of the coordinate system. If in the first coordinate system we have a set of numbers ϕ_1 and we then transform to a second coordinate system by a transformation

$$T_{21}: x_1^a \rightarrow x_2^a = x_2^a(x_1^a),$$

the set of numbers will then have a different form, say ϕ_2. Repeating the process to a third coordinate system and a third set of numbers ϕ_3,

$$T_{32}: x_2^a \rightarrow x_3^a,$$

$$\phi_2 \rightarrow \phi_3,$$

it is clear that so long as the numbers belong only to the physical system and not to the coordinate system it cannot matter whether we proceed straight from the first system to the third or go through the intermediate

† Section 3 entails rather more mathematical ability than the rest of the chapter, but a reader who is not particularly interested in quantum mechanics may proceed directly to section 5.

system. Using the following notation for the transformation of the numbers

$$t_{21}:\phi_1 \to \phi_2,$$

$$t_{32}:\phi_2 \to \phi_3,$$

$$t_{31}:\phi_1 \to \phi_3,$$

the principle of relativity then takes the form that

$$t_{31} = t_{32}\,t_{21}.$$

It is worth-while to make all this more precise, and to connect it with other developments, by introducing the idea of a group. A set of elements a, b, c, . . . said to form a group if there is a binary operation defined between the elements, say

$$a, b \to ab$$

(denoting the operation by juxtaposition as with multiplication in elementary algebra) under which the set is closed, i.e. ab always lies in the set when a and b do; and such that $a(bc) = (ab)c$. Further, there is assumed to exist a unit element I, such that $aI = Ia = a$, and every element a is supposed to have a reciprocal a^{-1}. We are here concerned with the group of all those transformations leaving $\eta_{ab}x^a x^b$ invariant, a set which the reader may easily verify to be a group according to the definition, if the binary opertion means applying the two transformations in succession. (This group has an obvious *sub-group*, those transformations leaving x^0 unchanged, which is the orthogonal group in three dimensions.) A requirement of the principle of relativity is then that the transformations of the sets of numbers form a *representation* of the transformation group. [By a representation R of a group G whose elements are g_1, g_2, . . . is meant an assigning of a *new* set of quantities g_i^R such that $g_i \to g_2^R$, where a binary operation is defined between the g_i^R as well, and has the property that $(g_i g_j)^R = g_i^R g_j^R$.]

Having formulated the principle of relativity, it is now necessary to look for representations of the coordinate group, which, in the case of special relativity, is the Lorentz group described above. It is in fact straightforward to find a whole set of representations of any coordinate group, known as the tensor representations as follows. One first decides on certain quantities which for physical, intuitive, reasons are obviously not properties of the coordinate system. One examines the method of transformation of these quantities and uses it to define the transformation of arbitrary sets of quantities in the same way. It is then certain that the transformations defined are a representation of the group. The simplest representation is the one in which the numbers are not changed at all by the coordinate transformation. This is known as the trivial representation, though the name is physically inappropriate. Quantities transforming under the trivial representation are the invariants of the group, or scalars.

The next step is to seek some quantities which will obviously not be properties of the coordinate system, and one set of quantities is that known already to mathematicians as the gradient of a scalar field

$$\frac{\partial \phi}{\partial x^a}, \quad (a = 0, 1, 2, 3).$$

Under any coordinate group of the form

$$x^a \rightarrow x^{a'} = x^{a'}(x^a)$$

the gradient quantities are known to transform in the following way:

$$\frac{\partial \phi}{\partial x^a} \rightarrow \frac{\partial \phi}{\partial x^{a'}} = \frac{\partial x^a}{\partial x^{a'}} \frac{\partial \phi}{\partial x^a},$$

which we may use to define the so-called representation of covariant vectors:

$$A_a \rightarrow A_{a'} = X^a_{a'} A_a,$$

where

$$X^a_{a'} = \frac{\partial x^a}{\partial x^{a'}}.$$

Another set of quantities that we can be sure are not properties of the coordinate system are the coordinate differentials $\mathrm{d}x^a$, which transform in the following way

$$\mathrm{d}x^a \rightarrow \mathrm{d}x^{a'} = X^{a'}_a \,\mathrm{d}x^a,$$

and so give rise to the representation known as the contravariant vectors

$$A^a \rightarrow A^{a'} = X^{a'}_a A^a.$$

In fact, for the case we are considering, that of the Lorentz group, there is a simple relation between the vectors, i.e. that if A_a is a covariant vector, then $\eta^{ab} A_b = A^a$ is a contravariant vector.

Having found the two types of vector representation we can now generate an infinite set of more complicated representations by taking products. For example, we could consider sets of quantities of the form

$A^a_b = B^a C_b$ where the suffixes on the vectors indicate the representation under which they transform, and we could then derive the representation of the form

$$A^a_b \to A^{a'}_{b'} = X^{a'}_a X^b_{b'} A^a_b.$$

These are the tensor representations, this particular one being a tensor of rank 2, since it has 2 suffixes.

This procedure has been explained at some length because, as will be seen in the next section, for historical reasons the generation of the tensor representations played an important part in the development of the subject.

4. The spin representations

The attempt to reformulate physics in a way consistent with special relativity was so successfully carried out from 1905 onwards by means of the tensor calculus that many physicists were of the opinion, whether or not they stated it explicitly, that only the tensor representations of the Lorentz group were of importance. Many people indeed thought that only the tensor representations existed. This belief suffered a rude shock[21] in 1928 when Dirac[22] discovered his relativistically invariant equation of the electron, and this shock led to a considerable further development of the theory, and also to its greatest triumphs. Dirac's equation involved quantities transforming under different representations of the Lorentz group from the tensor representations, and we must now make clear how these representations arise. It is sufficient for our purpose to consider the orthogonal group in two dimensions, since such representations already exist at this stage, and, as we have seen, the orthogonal group is a sub-group of the Lorentz group, so that such representations must also appear in any considerations of relativistic invariance. The new kind of representation arises when we consider the transformation of a contravariant tensor of rank 2 in two dimensions. More explicitly we have a transformation of the form

$$A^{ab} \to A^{a'b'} = X^{a'}_a X^{b'}_b A^{ab},$$

where the coefficients of the transformation are determined by a rotation of the axes

$$x^{1'} = x^1 \cos\theta + x^2 \sin\theta,$$

$$x^{2'} = -x^1 \sin\theta + x^2 \cos\theta.$$

When we determine the transformation equations explicitly we get

$$A^{1'1'} = A^{11}\cos^2\theta + (A^{21} + A^{12})\cos\theta\sin\theta + A^{22}\sin^2\theta,$$

$$A^{1'2'} = -\cos\theta\sin\theta\, A^{11} + \cos^2\theta\, A^{12} - \sin^2\theta\, A^{21} + \sin\theta\cos\theta\, A^{22},$$

$$A^{2'2'} = A^{11}\sin^2\theta - (A^{12} + A^{21})\cos\theta\sin\theta + A^{22}\cos^2\theta,$$

from which a number of conclusions follow.

In the first place any tensor of rank 2 may be written in the form

$$A^{ab} = \tfrac{1}{2}(A^{ab} + A^{ba}) + \tfrac{1}{2}(A^{ab} - A^{ba})$$

as the sum of a symmetric and a skew-symmetric tensor. These two parts are then obviously transformed independently by the transformation, the antisymmetric part being in fact an invariant in two dimensions. Coming now to the symmetric part, we may write the transformations in the form

$$p' = p\cos^2\theta + q\sin^2\theta + r\sin 2\theta,$$

$$q' = p\sin^2\theta + q\cos^2\theta - r\sin 2\theta,$$

$$r' = r\cos 2\theta + \tfrac{1}{2}(q-p)\sin 2\theta,$$

where $$p = A^{11}, \quad q = A^{22}, \quad r = A^{12} + A^{21},$$

from which we see at once that there is another invariant, i.e.

$$p' + q' = p + q.$$

The remaining terms can now be written in the form

$$\tfrac{1}{2}(q' - p') = \tfrac{1}{2}(q-p)\cos 2\theta - r\sin 2\theta,$$

$$r' = r\cos 2\theta + \tfrac{1}{2}(q-p)\sin 2\theta.$$

A most singular consequence of this re-writing is that we have constructed from the tensor of rank 2 a set of two quantities whose transformation has exactly the form of a contravariant vector, with this exception, that whereas the vector would have transformed under an equation whose coefficients consisted of circular functions of the angle, here it is the double angle that enters. The fact that this transformation is a representation of the orthogonal group is evident from the way in which it has been derived, so that from the vector representation we have derived a new representation which is essentially that of the tensor of rank 2. But by re-writing this new representation with the *single* angle instead of the double we have a set of two quantities $[r, \frac{1}{2}(q\text{-}p)]$ which are now components of a vector, and by reversing the argument the original vector transformation will now have the form

$$\phi^{1\prime} = \phi^1 \cos\theta/2 + \phi^2 \sin\theta/2,$$
$$\phi^{2\prime} = -\phi^1 \sin\theta/2 + \phi^2 \cos\theta/2,$$

where the half-angle enters.

Since we have been dealing exclusively with representations of the orthogonal group it is obvious that we have here *some* sort of representation. However, this representation is of a different kind from the tensor representations, because if we increase the original angle of rotation by four right-angles we do not change the transformation of the coordinates at all, but we do change the new quantities, and in fact turn them into their negatives. Accordingly an increase by eight right-angles leaves the new quantities unchanged as well. This is expressed by saying that the representation under which the new quantities, the so-called *spinors*,[23] transform, is a two-valued representation of the group. Corresponding to each transformation of the original group there are two elements of the representation. When Dirac was confronted with the problem of re-writing quantum mechanics in a Lorentz invariant fashion he found himself obliged to use spinors in order to do it. He was then able to modify quantum mechanics in a way which was consistent with special relativity. One of the consequences of his modification was the prediction that to every particle, such as an electron, there corresponds also an antiparticle, in this case the positron, and in 1931 the positron was identified experimentally by Anderson.[24] Later developments in quantum mechanics have confirmed completely Dirac's use of special relativity, and provide one of the principal bodies of supporting experiment for the theory.

5. The theory of gravitation

It is time to turn to the problem of including mechanics, by which we mean the mechanics of particles, and the theory of gravitation, in the framework of the Lorentz transformation.

The problem falls into two parts: first, to determine a reformulation of the mechanics of a particle under given (or no) forces, which agrees with that of Newton's at low velocities; second, to discuss the introduction of gravitational forces into such a theory.

The problem of formulating particle mechanics so as to be consistent with the Lorentz transformation can be answered in the following manner. Since Newtonian mechanics is known to hold very well for a wide range of low velocities let us make the assumption that it holds exactly when the velocity is zero, i.e. in the so-called rest-frame. If we take this frame as the x' y' z' coordinate system, so that initially the velocity is v in the xyz system, Newton's law that (mass) × (acceleration) = force takes the form

$$m\frac{\mathrm{d}^2x'}{\mathrm{d}t'^2}=F'.$$

From the transformation equations, however,

$$\frac{\mathrm{d}x'}{\mathrm{d}t'}=\frac{(\mathrm{d}x/\mathrm{d}t)-v}{1-v(\mathrm{d}x/\mathrm{d}t)},$$

so that, noting that after differentiation has been completed v is to be taken as $\mathrm{d}x/\mathrm{d}t$, we have

$$\frac{\mathrm{d}^2x'}{\mathrm{d}t'^2}=\frac{\mathrm{d}t}{\mathrm{d}t'}\cdot\frac{\mathrm{d}^2x/\mathrm{d}t^2}{1-v^2}=\beta^3\frac{\mathrm{d}^2x}{\mathrm{d}t}.$$

However, it is more useful to rewrite this result in a different form; since

$$\frac{1}{\beta^2}=1-v^2,$$

it follows that

$$\frac{-2}{\beta^3}\frac{\mathrm{d}\beta}{\mathrm{d}t}=-2v\frac{\mathrm{d}v}{\mathrm{d}t},$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\beta v)=\beta\frac{\mathrm{d}v}{\mathrm{d}t}+\beta^3v^2\frac{\mathrm{d}v}{\mathrm{d}t}$$

$$=\beta^3\frac{\mathrm{d}v}{\mathrm{d}t}.$$

In all, then, the law of motion becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}(m\beta v)=F',$$

and if, as usual in mechanics, this law is used to *define* mass (by, for example, collision experiments) it follows that the measured masses will vary with velocity according to the law

$$m' = \beta m = \frac{m}{\sqrt{(1 - v^2)}}.$$

Such an increase of mass with velocity had been known long before relativity from such experiments as those of Kaufmann.[25]

When this velocity–mass relation is compared with Newtonian mechanics by seeking the next approximation to $m' = m$ (when $v \to 0$), we get

$$m' = m(1 + \tfrac{1}{2}v^2 + \ldots)$$

so that the increase of mass is approximately equal to the Newtonian kinetic energy. This led Einstein[26] to identify $m' - m$ as the exact expression for the kinetic energy in relativity and to suggest that m' might be identified as the total energy, m being an additive constant known as the rest-energy, which the particle had even at rest.

This hypothesis was successfully confirmed by experiment, in the sense that the energy, as so defined above, is indeed conserved in atomic interactions, even those involving radioactive disintegrations. The formula $E = mc^2$ for the rest-energy (derived from the above by reinstating the velocity of light c, so as to be free to use CGS units) has become one of the best-confirmed consequences of the theory. In carrying the reformulation of mechanics forward to include gravitation, however, many new difficulties arise. This problem occupied Einstein for 10 years.[27]

One source of difficulty was the extremely accurate character of Newtonian gravitation. In the situation in which the theory predicted only very small discrepancies there was very little experimental evidence to suggest how it should be altered so as to agree with special relativity. The final solution of the problem was reached by Einstein in 1915[19]. By that time he had realized that the obstacle to incorporating Newtonian gravitation was the excessive reliance on the inverse square law.

As we noticed above, it was already known to Galileo[13] that the gravitational field causes bodies at a particular point to suffer equal accelerations. Einstein saw that the appropriate way of formulating the theory was to incorporate this observation of Galileo's first. The way in which we could do this is to transform at any particular point to a freely falling coordinate system. This coordinate system is then one in which, at the point under consideration, there is no gravitational field. So long as we can re-write our mechanics in a form which does not depend on the coordinate system employed we shall be able to re-write the equations in the new coordinate system with no gravitational field, and deduce the results in a normal coordinate system by transforming back again.

Naturally such a technique is not enough to remove all problems of gravitation. The gravitational field can be eliminated at one point, and so the strength of the field that we need to consider is measured by the extent to which this elimination cannot be simultaneously carried out at a neighbouring point. (That is, essentially, the measure of the field is really the rate of change of what it was for Newton.) This residual field[28] now varies with position and its variation has to be determined by certain field equations. These equations in turn must be equivalent in the first approximation to the inverse square law. Einstein discovered these equations and in this way was able to reconstruct the known description of the orbits of the planets.

Space precludes a detailed description of these equations, but we ought to know one feature of them. Since they are consistent with Newtonian mechanics, in which the inverse square law has the form

$$\text{Force} = \frac{Gm_1 m_2}{r^2},$$

where G is the constant of gravitation, m_1, m_2 the masses, and r the distance between them, they must involve the constant G.[29] Moreover since the object of this formulation is to produce a theory consistent with special relativity the speed of light is bound to play an important part. Once these two facts have been realized we can see an important feature of the gravitational field as described in general relativity. Imagine, for simplicity, a spherically symmetric body of mass M. If we work out the dimensions of the quantity GM/c^2, in which we have, for convenience, reinstated the term c for the speed of light, to enable us to use ordinary (say CGS) units for all quantities, it turns out to be those of a length, so that associated with the gravitational field of the mass is some intrinsic length. This situation is not an unfamiliar one. When some intrinsic quantity occurs in a theory it usually indicates the place at which the theory breaks down.[30] For the Sun the critical length defined here is of the order of 1 km.

The existence of this critical length was realized from the beginnings of the theory, and certainly from the solution of the field of the Sun discovered by Schwarzschild.[31] In this solution one carries out a boundary-value problem of a kind familiar in physics. Outside the Sun one finds a solution that happens, just as in the Newtonian theory, to be the same as the gravitational field that the Sun would have if all its mass were concentrated at its centre. Inside the Sun a different field results because the equations are different, since we are there determining the field not in free space but in matter. Boundary conditions enable us to join the two fields up so as to have a solution in the whole of space. Of course the critical radius never enters, since the radius of the Sun is very much more than 1 km and by the time we reach a radius of that order we are dealing with the interior field, for which there is no such critical length. When we work out the detailed field of the Sun we find that the next approximation to the orbit of the planets produces slowly rotating ellipses, and that the ellipses corresponding to the planet Mercury will in fact be rotating about 43″ per century, in astonishing agreement with the observed value.[32]

This test of the theory is in many ways its greatest triumph. During the first ten years of its existence a good deal of importance was attached to two other tests, the bending of light in a gravitational field, and the red shifts of spectral lines in an intense gravitational field.[33] Although the predictions of the theory are quite unequivocal in these two matters, the observations are very difficult to carry out accurately, and the rotation of the orbit of Mercury is still one of the most satisfactory confirmations of the theory.

6. Cosmology

Because of the extreme accuracy of the Newtonian theory it was regarded as still difficult in the twenties and thirties to confront general relativity with experiment in any satisfactory way; for this reason interest waned in the theory. Instead of investigating the basic assumptions, most of the workers in the theory turned to the consideration of cosmology and of the effects of relativity there. It had already proved possible to construct some cosmological models in Newtonian physics, but the modern movement in cosmology cannot really be said to have got under way until Seeliger and Neumann [34] independently raised objections to the infinite Newtonian universe. Essentially their objection was that when the volume of a Newtonian distribution of matter tended to infinity the gravitational potential became non-existent as did the gravitational force. Earlier Olbers, in 1823, had calculated that if the universe were homogeneous and infinite the night sky would be as bright as the average surface brightness of a star, whereas in fact it is dark.[35] Difficulties of this kind in Newtonian cosmology give rise to considerable interest in general relativity, in the hopes that the new theory would avoid them.

Einstein himself[36] was able to construct, with the help of a small modification of his field equations, a model of the whole universe with uniform density in which the matter was at rest. This was a great advance on the Newtonian cosmology despite the disadvantage of having to modify the field equations, because it allowed the description of matter in self-equilibrium. This equilibrium was, however, unstable and if the Einstein universe is disturbed the matter in it begins to recede from its original position.[37] Moreover, de Sitter[38] found another cosmological solution in which there was no matter, but in which the geometry of the space corresponded to a uniform expansion. These two special cases were reconciled in 1922–4 by the construction of expanding models by Friedmann[39] and Lemaître. Such expanding models could start from the Einstein model and expand asymptotically to the de Sitter form, or they could expand and contract.

Not long afterwards Hubble,[40] surveying the evidence about the velocities of forty-six nebulae, put forward his well-known velocity–distance relation in which each distant nebula was supposed to have a velocity of recession proportional to the distance. He did this in 1929 in ignorance of the fact that there was a theoretical basis in relativistic cosmology for such an expansion.

So far the contribution of general relativity to cosmology had been one of considerable success, but now this success came to be regarded as somewhat of an embarrassment. The universe is a unique object, but the relativistic theories of cosmology do not provide a unique model but an infinite range of models, with different kinds of behaviour. It seems as if some other considerations are needed in order to choose one such model. More serious questions of a philosophical nature can also be urged against relativistic cosmology. In adopting relativity as it stands the cosmologist automatically made the assumption that the laws of physics would be substantially unchanged during the whole evolution of his model. If, however, one has an expanding universe, the matter in it must originally have been extremely compressed, and at very high densities and, possibly at temperatures very different from now; there is then no way of telling whether what we now recognize as physical laws would still hold or not. Arguments of this sort were put forward in 1948 by Bondi and Gold.[41] In order that one could argue freely about events a long time ago and a great distance away, when circumstances were very different, these authors hold that one must make some definite assumptions about physical laws, and they argued that the only reasonable assumption to make first was that the laws were the same then and there as here and now, *because circumstances were the same.*

In view of the expansion of the universe the authors are therefore committed to an additional creation of matter to make up the density of the material that is receding from us.[42] Hoyle later rendered this theory consistent with general relativity[43] by constructing field equations for it. For a number of years the steady-state theory, as it came to be known, was well supported by experiment, but recently this support seems to be lacking.[44] It is too early yet to say whether or not the steady-state theory will survive the critical experimental observations now being made, but whether or not it survives, the arguments of Bondi and Gold are as valid as ever. Unless one has an acceptable theory of what physical laws would be like in vastly different circumstances from at present, one cannot have a cosmological theory without an assumption that the universe has always been much the same as it is now.

7. New problems in general relativity

During the thirties, apart from the work on cosmology, little was done in general relativity except for the monumental effort of Einstein, Infeld, and Hoffmann.[45] The field equations in relativity are nonlinear equations, and it is found that this nonlinearity leads to the following unusual situation: the singularities in the solution of the field equations are interpreted as the sources of the field, as usual, and in view of the nonlinearity the motion of these singularities is determined *by the equations themselves*. Thus the equations of motion of the theory *follow from* the field equations rather than being able to be postulated independently as, for example, the force on a moving electron is postulated, in electromagnetic theory, independently of Maxwell's equations.

Since 1945, however, many new developments have arisen in general relativity. If any one feature is to be observed in these new developments it is that they have all sought to find those aspects of the theory that are quite inconsistent with Newtonian mechanics. Whenever Newton's theory applies to a problem in gravitation it applies extremely accurately, so that it is very difficult to do better. Workers since 1945 have sought to find those aspects of the theory in which Newton's gravitation does not apply at all. The first of these problems to be satisfactorily dealt with was that of gravitational waves. Because of the principle of equivalence, and the fact that the theory is invariant under accelerated coordinate transformations, the very definition of a gravitational wave is beset with some difficulty. The most obvious definition by analogy with water waves or electromagnetic waves tends to give a gravitational field that is simply derived by transforming space without field in a periodic manner. As a result of work by a number of authors the concept of a gravitational wave was clarified and the fact that an oscillating body transmits waves that carry away energy (and so diminish the mass of the body) was well established.[46] It even proved possible to design apparatus for the reception of gravitational radiation,[47] although up to the present it has not been possible to design a generator of gravitational waves with sufficient strength for this to be observed.

From 1945 onwards a great deal of effort has also been put into the reconstruction of general relativity in such a way that it might be consistent with quantum mechanics, that is, in incorporating the uncertainty principle in general relativity.[48] The necessity of doing this has now become much more acute for reasons that we must explain at some length. When we consider a gravitating rigid body it is obvious, from the fact that the gravitational force between two particles is always attractive, that gravitational forces cannot be the only ones acting. If we imagine a uniform sphere that, from a certain moment onwards, contains matter acted on only by gravitational forces, the matter would collapse towards the centre. If we make the assumption that there is no overtaking, so that one portion of matter originally outside another always remains outside it, the matter that was originally at a distance R from the centre forms a shell that is under the attraction of all the matter inside, that is of an amount $\frac{4}{3}\pi R^3 \rho$, where ρ is the original density. If there is no overtaking, this matter, at any later time, finds itself under an attraction $\frac{4}{3}\pi G R^3 \rho / r^2$ per unit mass. It is now a routine matter to find how long this matter takes to reach the centre; it turns out to be a time

$$t = \left(\frac{3}{32G\rho}\right)^{\frac{1}{2}},$$

and since this is independent of R all the matter reaches the centre together, in conformity with the assumption of no overtaking.

By using the value of the constant of gravitation it follows that a sphere of the density of water under gravitational attraction alone will collapse completely in about a quarter of an hour. Evidently in real matter other forces must always be present. Ordinary matter consists of elementary particles of various kinds, the actual mass of the matter being mainly due to the baryons or heavy particles, of which the best known are protons and neutrons. Since matter is neutral the baryons will be accompanied by an appropriate number of electrons, of much smaller mass. We can therefore disregard the electrons in judging the gravitational forces. It has been calculated[49] that a collection of 560 baryons has an equilibrium state of 10 atoms of iron of atomic weight 56 arranged in a certain crystal lattice. The reason that the atoms are those of iron is due to the nuclear forces; the fact that a particular crystal lattice is determined is caused by the chemical forces. When we increase the amount of matter, the gravitational pull towards the centre increases continually, and eventually dominates these other forces. For example, even when there are 56×10^{41} baryons the equilibrium state consists of a sphere of iron of radius about 8 km. If more and more mass is added, there will eventually come a point at which the chemical forces are completely overcome. This occurs when the central density is about 5×10^8 g cm^{-3} and the mass is about 1·2 of that of the sun.

The electrons at the centre have now been squeezed to such a small volume that they combine with the protons and generate neutrons. Imagine that even more matter is added. Eventually the gravitational forces will have increased so much that they will also swamp the nuclear forces. A second collapse point arises with a central density of about 10^{16} g cm^{-3} and a mass of about 0·7 of the Sun's. Any further addition of matter would give rise to further contraction, and there appears to be nothing left to prevent the catastrophic collapse discussed above.

There are, however, grave doubts about whether such a collapse is possible in general relativity, since a body that is collapsing continually must eventually shrink to a radius that is less than the critical length mentioned before. There are many acute theoretical difficulties at this point in the theory, and not least of these is the following. Once a body has passed the second crushing point it seems it must eventually have a radius less than its critical length, and if we imagine a sphere of mass M and radius R, to which a mass m is brought up from infinity, we see at once that the small mass loses an amount of potential energy GMm/R or, because of the relation $E = mc^2$ between mass and energy in relativity, loses an amount of mass GMm/Rc^2. The amount added to the sphere is then not the amount that leaves the source of mass at infinity, but a smaller amount

$$m\left(1 - \frac{GM}{Rc^2}\right).$$

At the critical radius the energy exactly cancels out the added mass. If we imagine the mass brought up in the form of baryons this cancelling out is very difficult to understand, because according to quantum mechanics the number of baryons is always conserved, whereas here we are confronted with a process involving the systematic destruction of them. This apparent inconsistency points to the need of rendering relativity and quantum mechanics consistent with one another.

8. Recent observations

A great motivation behind the consideration of collapsing bodies has come during the more recent investigations of the most distant parts of the universe. Between 1945 and 1960 radio astronomers found a number of objects emitting radio signals but not correlated with optically observed objects. In 1960 Sandage found a 16th magnitude object which was a strong radio emitter, and also an optical star. But the spectrum of this object (3C48) was very unusual. It contained emission lines that did not agree with any of the common elements. Moreover, Sandage found that its brightness varied so that it could hardly be a very distant galaxy appearing to be a point only because of its distance, but must be a compact body such as a star. Later, Schmidt investigated 3C273, a similar object with emission lines in the wrong place and noticed that the four striking lines in the spectrum were in fact four hydrogen lines, red-shifted by 16 per cent, corresponding to a velocity of recession round about one-tenth of that of light. Greenstein and Matthews were now able to see that the spectrum of 3C48 was in fact that of hydrogen, red-shifted by an amount corresponding to a recession velocity about three times that mentioned above. It is too soon to pronounce definitely about such objects,[50] but at present there seems a balance of opinion in favour of the red-shift in the spectrum being due to a velocity of recession, this velocity being caused by the expansion of the universe, as with the nearer galaxies; a correspondingly very large source of energy is then needed to provide for the necessary radiation for such a very distant object to be visible. Hoyle and Fowler[51] suggested that whereas the normal thermonuclear process in stars would be quite inadequate to produce sufficient energy for these bodies, a source of energy that could be tapped was the potential energy of the gravitational field, which was released in collapse. Whether or not collapsing bodies are able to radiate in this way has still, however, to be determined.

Notes and References

1. Galileo, *Dialogue concerning two chief world-systems*, perhaps his most important work. Though it was published in 1632 he had begun the writing in 1626 and had discussed the contents with members of the Church, including a future Pope, as early as 1624.
2. In its clearest form in the laws of motion at the beginning of the *Principia*, London, 1686. See A. Motte's translation revised by F. Cajori, Cambridge University Press, 1934.
3. L. Euler in a paper written in 1739 but not published by the St. Petersburg Academy until 1750.
4. J. Bradley (1728) *Phil. Trans. R. Soc.* **35**, 637.
5. H. L. Fizeau (1859) *Annls Chim. Phys.* **57**, 385.
6. Fresnel's theory is contained in a letter to Arago ,who had studied the refraction of light from a telescope and showed that the amount of refraction is independent of the direction in which the telescope is pointed. This is in conflict with the notion of light as a wave motion in a medium through which the earth is moving. Fresnel's explanation leans heavily on the acoustical analogy.
7. The original experiment, in which a beam of light is split at a half-silvered mirror so that the halves traverse two paths at right angles and return, to interfere, was carried out by A. A. Michelson (1881) *Am. J. Sci.* **22**, 20 and repeated with considerable technical refinements by himself and E. W. Morley (1887) *Am. J. Sci.* **34**, 333. Later experimenters achieved like results, with the exception of D. C. Miller (1925) *Proc. natn. Acad. Sci. U.S.A.* **9**, 306. The general opinion now is that Miller's results are caused by an unknown residual error; for a full discussion see R. S. Shankland, S. W. McCuskey, F. C. Leone, and G. Kuerti (1955) *Rev. mod. Phys.* **27**, 167.
8. Maxwell's ideas come in a complete form in his *Treatise on electricity and magnetism*, Oxford University Press, 1873, which was therefore published two years after his election to the Cavendish chair in Cambridge. But he had been working on the ideas for much longer, and had made a rough sketch of the book during his time at King's College, London (1860–5).
9. W. Weber and R. Kohlrausch (1856) *Annln Phys.* **94**, 10 determined the ratio of electrostatic to electromagnetic units by comparing the capacity of a Leyden jar, measured by an electrostatic method, with the value calculated from a current discharge from the jar. Riemann's tentative proposals occur in a note of 1853 and a paper of 1858 submitted to the Göttingen Academy, but these were not published till after his death.
10. i.e. that the induced current always opposed the field producing it. *Annln Phys.* **31**, 483 (1834).
11. For example, F. E. Neumann *Die Prinzipien der Galilei–Newton'schen Theorie*, Leipzig, 1870, discusses the way in which the absolute space of Newton is replaced by the whole set of inertial frames, and uses the name 'the body Alpha' for the physical object specifying these frames.
12. The first paper was that of H. A. Lorentz, (1904) *Proc. Acad. Sci. Amst.* **6**, 809; an attempt to clarify and extend the ideas was in H. Poincaré (1906) *Rc. Circ. Mat. Palermo* **21**, 129 (submitted July, published December 1905). Their claims, against that of Einstein, have been championed by Sir Edmund Whittaker, in the chapter 'The relativity theory of Poincaré and Lorentz' in his *A history of the theories of aether and electricity: the modern theories* (1900–26), Nelson, London, 1953; but his judgement of the relative importance of these papers and

Einstein's must be reckoned to be that of a mathematician rather than a physicist.

13. Galileo, *Dialogue concerning two new sciences*. Completed 1636 and published Amsterdam 1638.
14. A. Einstein (1905) *Annln Phys.* **17**, 891, a paper that astonishes by its freshness. It is apparently written in ignorance of those mentioned in note 12.
15. The systematic exploitation of this method to derive, and to some extent generalize, special relativity is essentially due to E. A. Milne and may be found in a complete form in his *Kinematic relativity*, Clarendon Press, Oxford, 1948, though he developed it much earlier. The present popularity of it is due to H. Bondi.
16. The method given here, employing a common light-signal for two observers and one event, is due to Milne (see note 15). Einstein, in the paper mentioned in note 14, proceeded in a much more cumbersome and somewhat confusing manner.
17. L. W. Thomas (1927) *Phil. Mag.* (7) **3**, 1.
18. H. Minkowski (1908) *Göttinger Nachrichten* p. 53, gives some of the development: a clearer view is in his address to the 80th assembly of German physicists in Cologne, 21 September 1908, which is available (in translation) in *The principle of relativity* by A. Einstein and others, Dover, New York, 1923. In fact, Minkowski was partly anticipated by Poincaré (see note 12) but Poincaré's geometric formulation was insufficient to deal with the electro-magnetic field.
19. In his paper on the general theory of relativity, *Annln Phys.* **49**, 769 (1916).
20. See, for example, *The classical groups*, Princeton, 1939 and, of course, his very beautiful later work, *Symmetry*, Princeton, 1952.
21. For example, C. G. Darwin (1928) *Proc. R. Soc.* A **118**, 654. Darwin's view is quoted as being a great influence by Sir A. S. Eddington, *Relativity theory of protons and electrons*, Cambridge University Press, 1936.
22. P. A. M. Dirac (1928) *Proc. R. Soc.* A **117**, 610 or in his *Principles of quantum mechanics*, Clarendon Press, Oxford, 1930.
23. A full bibliography is given in E. M. Corson (1953) *Introduction to tensors, spinors and relativistic wave-equations*, Blackie, London.
24. C. D. Anderson (1932) *Phys Rev.* **41**, 405. Though, in fact, Anderson was working empirically, in complete ignorance of Dirac's theoretical investigation, and was inclined at first to identify the positively-charged particles as protons. It ultimately proved impossible to make the masses agree.
25. W. Kaufmann (1902) *Göttinger Nachrichten* **143**, describes his experiments in which he concluded that the measured mass of the electron at high velocities exceeds that at low velocities.
26. This expression for the kinetic energy occurs already in the paper of note 14, but the deduction of it there is somewhat confusing. In fact a new physical hypothesis is being made, although one for which the theory provides strong suggestions, viz. that the total energy, as defined here, is conserved in such reactions as radioactive disintegrations. For further details see H. A. Bethe, *Rev. Mod Phys.* **9**, 69 (1937).
27. A 'progress report' is his paper, *Annln Phys.* **35**, 898 (1911).
28. The means of describing this field are to be found in T. Levi-Civita (1926) *Math. Annln* **97**, 291.
29. The value of G, which was originally determined by H. Cavendish, *Phil. Trans. R. Soc.* 1798, p. 469, using an apparatus designed and constructed years before by the Rev. J. Michell, is $6{\cdot}67 \times 10^{-8}$ CGS units.
30. Perhaps the most striking examples are the dimensionless numbers in hydrodynamics, e.g. Reynolds' number, $\rho\frac{VL,}{\nu}$ where V is the fluid velocity ρ· the

density, L a typical length, and the ν coefficient of viscosity. This measures the relative value of inertial and viscous forces. See, e.g. G. Birkhoff, *Hydrodynamics*, Princeton, 1950.

31. K. Schwarzschild 1916, *Sber. preuss. Akad. Wiss.*, p. 189, gave the field of a mass-point, but he extended it to a spherical body later in the same year (p. 424).
32. To be more correct, the elliptical orbit actually rotates per century by an amount 5600″, of which a Newtonian calculation of the perturbation due to the other planets accounts for 5557″, leaving 43″ to be explained. For further details see G. M. Clemence (1947) *Rev. mod. Phys.* **19**, 361.
33. For a description of these and other experimental tests of the theory see *Gravitation* (edited by L. Witten), Wiley, New York, 1962.
34. H .Seeliger (1896) *Astr. Nachr.* **137**, 129 and C. Neumann (1896) *Allgemeine Untersuchungen über das Newtonsche Prinzip der Fernwirkungen*, Leipzig, 1896. In the discussion of cosmology I have derived great benefit from J. D. North, *The measure of the universe*, Clarendon Press, Oxford, 1965.
35. We mainly owe the wide knowledge of Olbers's work to H. Bondi, *Cosmology*, 2nd edn, Cambridge University Press, 1960. The paradox does in fact occur as early as the eighteenth century, being fairly fully expressed by E. Halley (1720) *Phil. Trans. R. Soc.* **31**, 23, and in the appendix to P. L. de Chéseaux, *Traité de la Comète qui a paru en décembre* 1743, Paris, 1744.
36. A. Einstein (1917) *Sber. preuss. Akad. Wiss.* **142**.
37. Or so it is usually stated in the books. However, W. B. Bonner (1954) *Z. Astrophys* **35**, 10, showed that such an instability results only for disturbances causing a discontinuity in the pressure.
38. W. de Sitter (1917) *Proc. Acad. Amst.* **19**, 1217; see also *Mon. Not. R. astr. Soc.* **78**, 3 (1917).
39. A. Friedmann (1922) *Z. Phys.* **10**, 377; see also *Z. Phys.* **21**, 326 (1924). These papers were largely ignored, and the work done again by both G. Lemaître (1927) *Annls Soc. scient. Brux.* **47**, 49 (translated in *Mon. Not. R. astr. Soc.* **91**, 483 (1931)) and H. P. Robertson (1928) *Phil. Mag.* **5**, 835.
40. C. Wirtz had tentatively suggested that a systematic effect in nebular spectra, the so-called *K-term*, might be caused by recession from the Sun (*Astr. Nachr.* **216**, 451 (1922)) but the full law is Hubble's; *Proc. natn. Acad. Sci. U.S.A.* **15**, 168 (1929). He seems to have found it in ignorance of the theories of Friedmann and Lemaître, which is supported; strikingly in analogy with Anderson confirming the existence of the Dirac positron in ignorance of the theoretical work. It is striking that Hubble's generalization involved the knowledge of velocities of forty-six objects, but of *distances* of only eighteen of these.
41. H. Bondi and T. Gold (1948) *Mon. Not. R. astr. Soc.* **108**, 252.
42. Not that Bondi and Gold were the first authors to contemplate continual creation of one kind or another; the idea goes back to W. D. MacMillan (1918) *Astrophys. J.* **48**, 35. It is only with Bondi and Gold, however, that the philosophical problem mentioned in the text is seriously considered.
43. F. Hoyle (1948) *Mon. Not. R. astr. Soc.* **108**, 372. Hoyle's paper was published later than Bondi and Gold's but written at about the same time.
44. It would be inappropriate to discuss these very recent experimental results here; essentially they arise from observations of radio-astronomy that appear to imply a greater density of certain peculiar bodies in more distant regions. Such a difference, consistent with a single creation if these bodies are in fact very old, would be quite inconsistent with the steady-state theory.
45. A. Einstein, L. Infeld, and B. Hoffmann (1938) *Ann. Math.* **39**, 65. Their methods were later much simplified by many workers. A full description, with biblio-

graphy, is to be found in L. Infeld and J. Plebanski *Motion and relativity*, Pergamon Press, London, 1960. The equations of motion were found independently by V. Fock (1939) *Zh. eksp. teor. Fiz.* **9**, 375, and many subsequent papers, summarized in his book *The theory of space, time and gravitation* (translation by N. Kemmer), Pergamon Press, London, 1959.

46. The most valuable of a number of contributions are to be found amongst H. Bondi, F. A. E. Pirani, and I. Robinson (1959) *Proc. R. Soc.* **251**, 519; H. Bondi (1960) *Nature, Lond.* **186**, 535; I. Robinson and A. Trautmann (1962) *Proc. R. Soc.* A **265**, 463 (1962); H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, *Proc. R. Soc.* A **269**, 21 (1962). But see also the work on classification of radiation fields, beginning with the emphasis on Petrov's classification by F. A. E. Pirani, *Phys. Rev.* **105**, 1089 (1957), and continuing also with the theory of rays of R. K. Sachs, *Proc. R. Soc.* A **264**, 309 (1961) and **270**, 103 (1962).
47. J. Weber (1960) *Phys. Rev.* **117**, 306.
48. For an able summary of this effort up to 1960 see the chapter 'Status report on the quantization of the gravitational field' by P. G. Bergmann and A. B. Komar, in *Recent developments in general relativity*, Pergamon Press, London, 1962. Since then no great advances have been reported.
49. Much of the emphasis on the importance of the collapse phenomenon is due to J. A. Wheeler, from whom the following section is taken. The most complete recent summary of his views and those of his school is to be found in B. K. Harrison, K. S. Thorne, M. Wakano, and J. A. Wheeler, *Gravitation theory and gravitational collapse*, University of Chicago Press, 1965.
50. Although it is early to assess the observations, reference may be made to G. R. Burbidge, E. M. Burbidge, and A. R. Sandage (1963) *Astrophys. J.* **137**, 1005 and J. L. Greenstein and M. Schmidt (1964) *Astrophys. J.* **140**, 1.
51. F. Hoyle and W. A. Fowler (1963) *Mon. Not. R. astr. Soc.* **125**, 169 and *Nature, Lond.* **197**, 533 (1963).

3 *Matter and Radiation*

Introduction

It is convenient to take 1895, the year in which Röntgen discovered X-rays, as the starting-date of our period. Not only was his discovery important in itself, though almost an accidental one, but it led directly to the discovery of radioactivity, and hence of radium and the whole of nuclear physics. After a curious delay, it led some twenty years later to the first exact knowledge of how atoms are arranged in solids; and this has developed into the solid-state physics of the present day, which because of its wide commercial applications occupies the attention of about half the physicists now working.

Ideas in physics are more often forced on physicists by unexpected experimental discoveries than invented to explain long-known facts. Even relativity (see Chapter 2) (which is perhaps better regarded as the final expression of nineteenth-century physics than as part of a new age) would not have been accepted without the experimental evidence against a stagnant ether.

The outstanding characteristics of this new age of physics are two: first a great increase in the importance of the idea of 'atomicity', namely that the fundamental things in the universe, and many of the less fundamental, are mass-produced as vast numbers of identical units; second, the increasing realization that the act of making an observation necessarily alters that which is observed, and that on an atomic scale, and still more on a sub-atomic scale, this alteration may be so important and so unavoidable as to lead to results that seem contrary to common sense and require a revision of very fundamental ideas.

Let us begin with the first of these. The idea of atoms, which is as old as the Greeks, had been adopted by the chemists since Dalton's work at the beginning of the nineteenth century. They had discovered elements of which some eighty-five were known in 1895, and vast numbers of their compounds, for each of which they knew the number of atoms of the different kinds concerned that formed the unit called the 'molecule', the smallest piece of matter having the chemical and other properties of the compound. For many compounds, particularly among the numerous and complicated compounds of organic chemistry, they also knew which atoms were directly tied to which inside the molecule. They even knew the approximate shapes of some of these molecules.

Physicists had done less with atoms and molecules. Some even disliked the idea because atoms had never been directly observed, but they were a minority whose pedantry did little harm. Molecules were indeed a fundamental idea in the kinetic theories of Boltzmann and Maxwell that were an established part of the physics of gases. Otherwise, physics dealt mostly in continuous media, both material and etherial, which were supposed to be divisible into infinitesimal parts without altering their properties.

The great discoveries of the twentieth century have shown that atoms have indeed a structure, and are composed of not more than three basic units now called electrons, protons, and neutrons. The structure is a peculiar one in that all the mass except about 1 part in 4000 is concentrated in a small central body of radius of the order of 1/10000 that of the atom. The rest of the atom is made up of the very light electrons and it is they, by virtue of their arrangement, that give an atom nearly all its characteristic properties, except mass and radioactivity. Since the last war a number of other bodies have been discovered, that are called, for want of a better name, 'fundamental particles'. All of them are very ephemeral, the longest-lived lasting only a few millionths of a second. Their masses vary, but none measured so far differs from that of a proton by as much as a factor of 10 either way. Their relation to one another and to the more stable basic units is the most exciting problem of present-day physics.

The problems associated with the philosophy of observation are most acute for the lightest known body, the electron. The study of these problems has led to an understanding of the distribution of the electrons in atoms explaining, among other things, the chemical differences between the atoms of different elements and the linking together of atoms to form the molecules of compounds. Chemistry indeed has become a branch of physics in the sense that there is good reason to believe that the structures of all compounds and the reactions between them are describable in terms of the mathematical laws that relate to electrons. The mathematics is very complicated in most cases, perhaps virtually insoluble in some, but quite a number have been studied with success.

The existing state of physics

One cannot understand the advances, and particularly the new ideas, of the period without some knowledge of the position from which they started.

Newtonian mechanics, the consequence of Newton's three laws of motion, which make mass a fundamental concept, was the apparently firm foundation on which all physics rested. This was so taken for granted that 'explanation' meant explanation in these terms. To it was added Newton's law of universal gravitation. Much of everyday physics—the elastic behaviour of solids, the viscous flow of liquids, the expansion and elasticity of gases—fitted the framework well as long as one did not expect to be told why the various 'constants' that appeared in these laws had the values for any particular substance that experiments showed they did.

By the middle of the nineteenth century heat was well established as being the random movement of the small particles of which bodies are made; oscillatory in solids, rectilinear paths interrupted by frequent collisions in gases, and presumably something between for liquids.

The laws of electricity, both of static charges and currents with their accompanying magnetic fields and a rather *ad hoc* knowledge of the magnetic properties of iron and other magnetic substances, were well established. Hertz had confirmed by experiment the form of these laws advocated by Clerk Maxwell by producing in the laboratory the 'wireless' waves that Maxwell had foretold.

For nearly the whole century it had been regarded as proved that light was a form of wave motion in an ether, and Newton's idea of a curious mixture of waves and particles was quite discarded. Efforts to find a satisfactory mechanical ether had not been too successful but this did not seem to matter so much since Maxwell's theory obviously proved light to be an electromagnetic effect, thus needing an electromagnetic ether, which it was thought might be found.

White light, as Newton had shown, can be analysed by a prism into its constituent colours. Under suitable conditions most elements can be made to emit light, also coloured, which when similarly analysed shows, not a continuous band of colour, but a group of discrete colours called a 'line spectrum'. The wavelengths of the various lines can be measured and the group of numbers so found forms a criterion that acts like a fingerprint to prove the presence of the element in the flame or electric discharge giving the light.

Attention was focused on the 'medium', which in some cases would be the ether, in others the surrounding matter. Thus even the electrical charges on material bodies, though they figured prominently in the mathematics, were surmised by Maxwell to be merely the way in which stresses in the medium became observable.

The idea of energy as something that could appear in many forms and be transferred from one piece of matter to another in the form of radiation or by contact, but could never be created or destroyed, was very important in the latter half of the nineteenth century. This law of conservation of energy and the allied one of the limited availability of energy, which forms the second law of thermodynamics, deny the possibility of perpetual motion. They allow of general arguments that led to remarkably far-reaching and fundamental conclusions in many branches of physics and chemistry.

X-rays

On the evening of Friday 8 November 1895 Röntgen found that on passing an electric current through a highly rarified gas contained in a glass vessel, a sensitive screen placed on the table fluoresced. This alone would not have been surprising, since the screen was there to detect radiation, such as the ultra-violet radiation known to be produced in such a discharge tube. The surprising thing was that the glass tube was covered with black paper that would certainly have stopped all radiations then known. Not only were the rays able to go through black paper but through wood and flesh, which were more transparent to them than was glass; quite a part of these produced must have been absorbed in the walls of the vessel. The absorption of the rays seemed to depend only on the density of the absorber. (This is not strictly true but a good first approximation.) The rays were not deflected by a magnet. They were weakly reflected but not refracted. They came from the region where the cathode rays hit the glass wall (the nature of cathode rays was still in dispute). Soon after his first paper, which appeared on 28 December 1895, Röntgen found that they made the air a conductor of electricity.

Ordinary air is a good insulator unless the voltage applied is enough to produce a spark or glow.

J. J. Thomson and his pupil Ernest Rutherford, then in his second year of research at Cambridge, studied this conductivity. They showed that, like the conductivity of aqueous solutions of salts, it is due to minute electrically charged particles, positive and negative, probably not much, if at all, larger than molecules. Those in solution were called *ions* and the same name was applied in the gases. However, in solutions the ions occur spontaneously and when carried away by the current are spontaneously replaced, whereas those in gases have to be made by the X-rays (or some other cause) and when made disappear in a few seconds, the attraction of the electric charges bringing ions of opposite sign together, so that they neutralize each other and can no longer be observed. An early X-ray tube is illustrated in Pl. 1.

Cathode rays

These had been discovered in the middle of the nineteenth century, soon after improvements in air pumps made it possible to get vacua in glass apparatus such that only about one part in a million of the initial air, or other gas, remained. In such conditions a potential of a few tens of kilovolts produces a discharge between metal electrodes sealed into the glass; this discharge shows, among other colour effects, a beam of light coming from the negative electrode (cathode), which when it reaches the wall of the tube makes the glass fluoresce with a colour that depends on the kind of glass. The rays can be deflected by a magnet held near the discharge. Hertz had proved that they could go through thin gold foils. Lenard examined the properties of the rays that escaped into the air through such a metallic film and showed that they were absorbed by matter in proportion to the density, the same weight per unit area of hydrogen or gold producing approximately equal absorption. In 1895 Perrin showed that the rays had a negative charge† associated with them. Their nature was hotly disputed. Most of the German physicists thought that they were etherial waves of some special kind. The British and

† The terms *positive* and *negative* electricity go back to Benjamin Franklin's 'one-fluid' theory of electricity. They are appropriate in the sense that equal charges of opposite sign placed together cancel out, and moving in opposite directions are currents in the *same* direction, but Franklin had very little reason for which he called 'positive'; this was in 1895 pure convention.

French that they were negatively electrified particles; such particles if moving would constitute a current that would be deflected by a magnet as the cathode rays were seen to be. This was one argument for the particle view; another was Perrin's recent experiment. The main arguments against it were that the rays could penetrate thin solids without leaving holes and that Hertz had apparently shown that they were not deflected by an electric field, for example by passing between two parallel plates connected to the poles of a battery, as moving charged particles ought to be. J. J. Thomson in 1897 answered this second and most serious objection by actually producing such a deflection and using it to determine the ratio *e*/*m* of the charge *e* to the mass *m* of the assumed

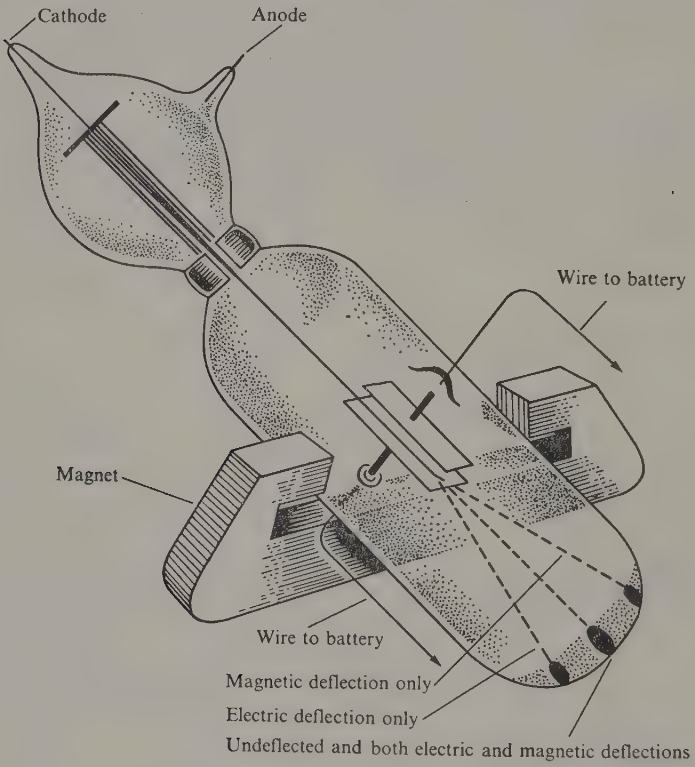


FIG. 3.1. J. J. Thomson's e/m experiment. The electric field between the plates is found from the voltage of the battery. The field of the magnet is measured in an auxiliary experiment. The velocity of the rays is found by balancing the two deflections.

particles (Fig. 3.1). The reason for Hertz's failure was that the residual gas in the tube became ionized by the rays and the ions moved close up to the parallel plates and neutralized the field. This could be avoided by improving the vacuum. It is probable that Thomson was helped to recognize this cause by his study of the ionization produced by X-rays in which similar effects occur. He was certainly influenced by an experiment of Goldstein's, which he quotes, but which was capable of other explanations. The value found for e/m was about 1000 times greater than that for the hydrogen ion in electrolysis, known even then with accuracy. Actually it is 1836 times greater. With considerable boldness, and influenced largely by Lenard's result quoted above, Thomson asserted that the particles of cathode rays, which do not depend on the nature of the gas or of the cathode used, form a universal constituent of matter. In the next two years he was able to prove that 'corpuscles' of the same e/m occur in two other effects, namely in the photoelectric effect, another discovery due to Hertz, and in the electrification produced by hot metals, a much-studied effect going back in the eighteenth century. For the photoelectric effect he was also able to measure the charge e, showing that it was the same as for an ion produced by X-rays, which in turn was equal to that on a hydrogen ion in electrolysis to the accuracy to which this last quantity was then known. Shortly afterwards, Townsend, also working at the Cavendish Laboratory in Cambridge, was able to make an accurate comparison between the last two charges by an ingenious method that avoided a direct measure of either but proved their equality.

As a result of these and other experiments, Thomson's thesis of the universality of the new particles was generally accepted in the early years of the new century, but the name 'corpuscle' was replaced by 'electron'. It had been clear from the beginning that the electrons could not be the only constituent of atoms, which in their normal state are electrically neutral, and it was improbable that they even constituted any large fraction of the mass. By about 1906 rather inaccurate but consistent calculations based on widely different effects made it probable that the number of electrons in an atom was comparable with its chemical atomic weight, taking hydrogen as unity. Thus the electrons account for a negligible fraction of the mass.

Electrons were regarded as small charged spheres. It could be proved that a charge could not be concentrated into an indefinitely small region, assuming Maxwell's theory of electricity, because such a region would acquire infinite mass. For a particle of the measured mass of an electron the region within which the ordinary laws must break down is of the order of a hundred thousandth of the diameter of an atom.

Radioactivity

In February 1896 Henri Becquerel found that compounds of uranium were able to blacken a photographic plate wrapped in black paper. The experiment was made in an attempt to find some effect analogous to X-rays. It was then considered that X-rays were connected with the phenomenon of phosphorescence (which they are not), and uranium was chosen because some of its compounds show this effect; but Becquerel soon showed that the effect is an atomic one that occurs equally well for the non-phosphorescent compounds. Like X-rays the rays from uranium render a gas conducting.

Rutherford made use of this last effect, which is easier to make quantitative than the photographic one, and showed that the rays are of two quite different kinds which he called α- and β-rays, the former completely stopped by a few centimetres of air, the latter able to go through a millimetre or two of aluminium. The even more penetrating γ-rays were discovered by Villard; they are in fact identical with X-rays except for their method of production. The β-rays were soon proved to be electrons of varying energy, sometimes very great. The study of α-rays led Rutherford to some of the most important discoveries in physics.

It soon became plain that the rays from uranium were complex in a further sense, namely that some came from other elements mixed with the uranium. The discovery of polonium and of radium by Madame Curie showed that a large part of the activity of an ore of uranium is due to the presence of small quantities of these and other elements mixed with it. They could be separated from it chemically. The same is true of ores of thorium, another radioactive mineral. By a study of thorium Rutherford and Soddy showed that a large proportion of the activity was left behind when the thorium itself was precipitated from a solution. More surprisingly, the thorium examined over a period of 20 days recovered its lost activity while that left behind lost its activity simultaneously, the sum of the two activities remaining unchanged. From these and similar experiments Rutherford and Soddy (1902) put forward the very bold hypothesis that the radioactive atoms—in this case those of thorium—were breaking down to make other radioactive atoms, those of thorium X, as they called the substance left in the solution, while these broke down in their turn giving off their own radiation in the process. The curves of decay and recovery are exponential, which implies mathematically that the number of the X atoms breaking up per second is proportional to the number present and the number produced per second from the mass of thorium is constant (Fig. 3.2). Since thorium is derived from minerals hundreds of millions of years old, this last is what one would expect, but the idea that thorium atoms can spontaneously emit radiations and at the same time produce an unstable atom ThX, which breaks down in about 20 minutes producing more radiation, was shattering to any conservative chemist used to the indestructability of atoms.

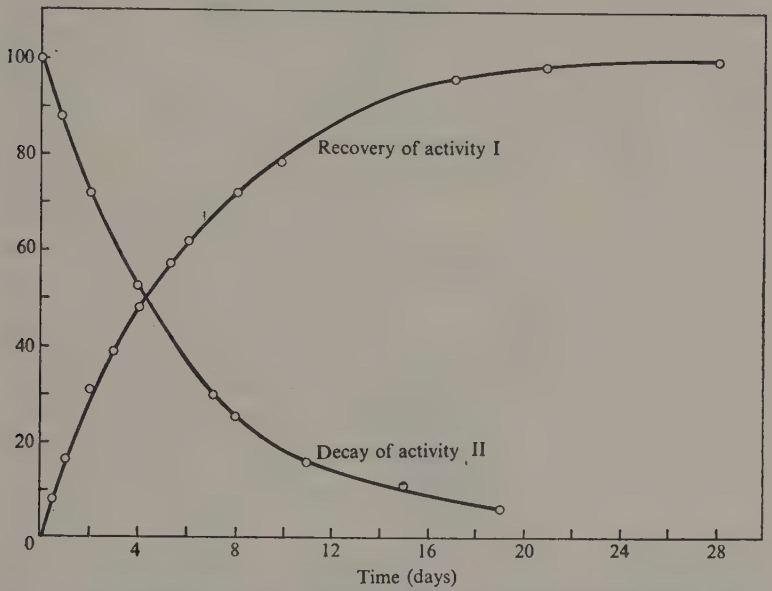


FIG. 3.2. Decay and growth of thorium X. Allowance has been made in this figure for initial irregularities in the first 2 days, due to short-life products. From E. Rutherford, *Radioactivity* (Cambridge University Press, 1904).

It was bad enough that it should be possible to remove electrons from atoms by X-rays and other means, but at least they eventually went back when the ions recombined. This new effect if true at all, must be a permanent change, and if the theory were to hold generally for all radioactive bodies it must apply to radium, an element quite unremarkable chemically that fitted nicely in the periodic table and whose atomic weight could be measured. However, support for the theory of successive disintegration accumulated as more and more radioactive substances were discovered that were ultimately derived from uranium or thorium and present in their ores but many of which could be produced in the laboratory by decay of others. In no case could the rate of decay or the emission of radiation be altered by chemical change or temperature.

The most conclusive proof came from the study of α-rays by Rutherford and his school. He found that they carried a positive charge and could be deflected, with some difficulty because of their great energy, by electric and magnetic fields (Fig. 3.3). By a process similar in principle to that used by Thomson for cathode rays, Rutherford showed that *e*/*m*, the charge-to-mass ratio for the α-particles, corresponds to a mass equal to two hydrogen atoms joined with a unit positive charge or to one helium atom with two charges. Two hydrogen atoms in fact make a molecule, but enough was known by then of the behaviour of ions in gases to make it most unlikely that a positively charged molecule of hydrogen would survive the path of destruction that the α-particle made through the air for anything like the several centimetres observed. Rutherford plumped for helium. This light inert gas was originally discovered by its spectrum

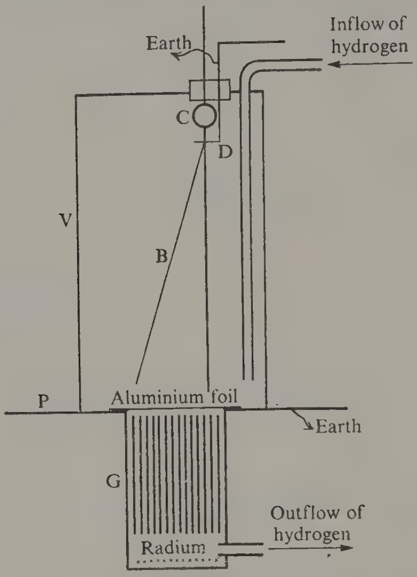


FIG. 3.3. Deflection of α-rays by a magnetic field (earliest method). The α-rays pass through narrow slits through brass plates into the electroscope, which detects them. When a strong magnetic field is set up perpendicular to the diagram the paths of the rays are curved and none get through. From E. Rutherford, *Radioactivity* (Cambridge University Press, 1904).

in the sun but had been found on earth included in some minerals, notably those of uranium and thorium, in itself a strong hint. Among the sources of α-rays was a heavy inert gas, radium emanation (now 'radon'), which appeared spontaneously in specimens of radium. Ramsey and Soddy were able to show in 1903 that the gases coming from a solution of a radium salt gave a spectrum containing unknown lines, which they attributed to the emanation but *not* at first the well-known helium lines. These, however, all appeared after 4 days, by which time the emanation would have lost most of its activity. The final proof that this helium was indeed the actual matter of the α-rays came three years later when Rutherford's assistant, Baumbach, was able to put the emanation into glass tubes so thin that the α-rays could pass through the walls. On placing such a tube in an evacuated vessel for a few days the spectrum of helium was detected in the vessel.

The changes of radioactivity are marked by a great release of energy, which if all the rays are absorbed appears as heat. This production of heat continues unceasingly and has done so in the case of uranium for thousands of millions of years. Rough calculations showed quite early that the energy of the process, wherever it came from, was of the order of a million times greater per unit mass than that of an ordinary chemical reaction. Shielding by thick blocks of lead had no effect on the heat produced, which made it unlikely that the heat is due to radiation from outer space. It was clear, and became gradually accepted, that the atom of uranium breaks up spontaneously emitting an α-particle, that the product breaks up in its turn and so on, each step emitting either an α- or a β-particle, sometimes accompanied by γ-rays. There are fourteen stages in all in the uranium series, not counting alternative paths, ending in an inactive form of lead. Thus atoms, at least certain atoms, break up spontaneously with the release of vast amounts of energy. Transmutation of the elements was a proved fact, but so far limited to a few heavy atoms, and uncontrolled.

The next major advance came from realizing the importance of a small but surprising effect. One of Rutherford's pupils, Marsden, found that a small proportion of α-rays incident on a film of gold, thin enough for most of them to go straight through, were so much deflected that they came back out of the foil on the side from which they had entered (Fig. 3.4). In an experiment with platinum, 1 in 8000 were reflected from a thick target. Rutherford afterwards said 'I remember . . . Geiger coming to me in great excitement and saying "we have been able to get some of the alpha particles coming backwards". It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you.' It was not difficult to show, on general mechanical grounds, that the α-rays, which have the mass of helium atoms, must have been in effective collision with something much more massive than an electron and something that produces a strong electric field. Geiger and Marsden of course continued the research and made accurate measurement of the effect. This was in 1909. In 1911 Rutherford hit on the explanation that a central charge provides the strong electric force and is associated with most of the atomic mass to minimize its recoil (Fig. 3.5). In fact about 100 units of charge were needed for gold. This was the origin of the 'nuclear' atom, though the word was not used in the original paper. It was an enormous advance, though some years before Lenard had suggested an atom mostly empty but with a few centres of force. Though the experiments were unable to determine the sign of the charged nucleus, the probability was greatly in favour of its being positive to neutralize the negative electrons, especially as the magnitude of the charge came out at about half the atomic weight, in general agreement with the number of electrons indicated by other methods. Later it was proved that the charge,

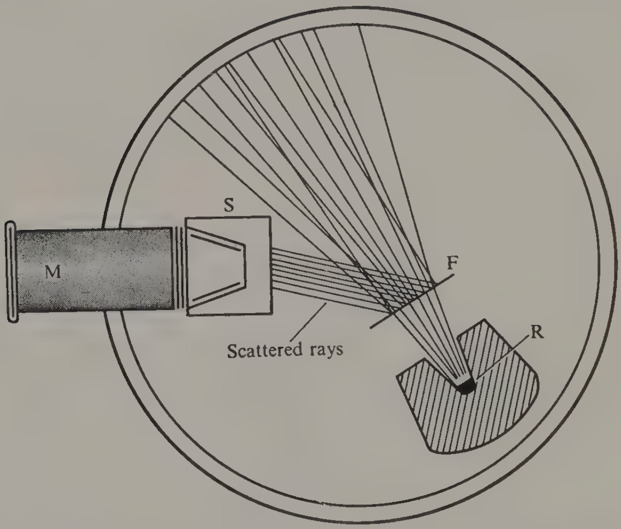


FIG. 3.4. Scattering of α-rays: Geiger and Marsden. The box, shown with its lid removed, turns on an air-tight joint carrying the microscope *M* and the screens, round the foil *F*, which is irradiated by the source *R*. For clarity the only scattered rays shown are those that hit the screen *S*.

and consequently the number of electrons round it, is equal to the 'atomic number', the rank of the element in a series in which the elements are arranged in ascending order of atomic weights.

In his first paper Rutherford pointed out that the great energy of the α-rays could be partly explained if they came from the nucleus by the mutual repulsion of the two positive charges. In fact, it was correctly assumed that all radioactivity was a nuclear effect.

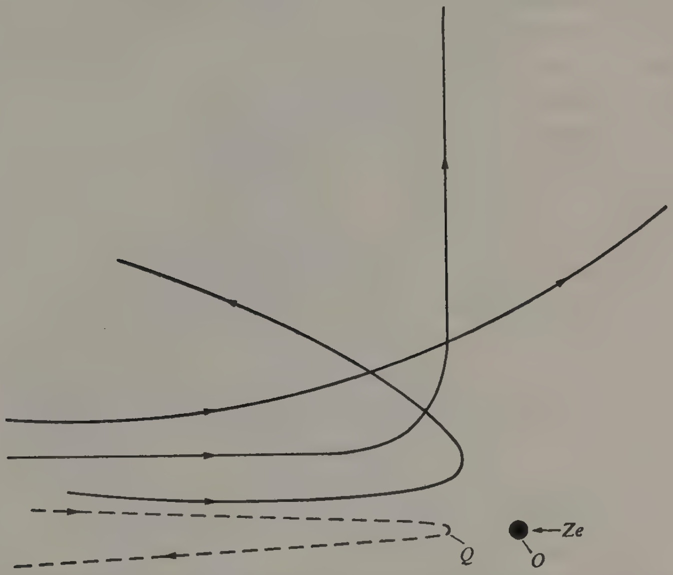


FIG. 3.5. Paths of α-rays scattered by a nucleus of charge *Ze*. The paths of the α-rays are hyperbolas except for those straight on the target, which go and return (dotted lines). Redrawn from M. Born, *Atomic physics* (Blackie, 1944).

This seems to be the point at which radioactivity passes over into the newer and wider conception of nuclear physics, which we shall discuss later.

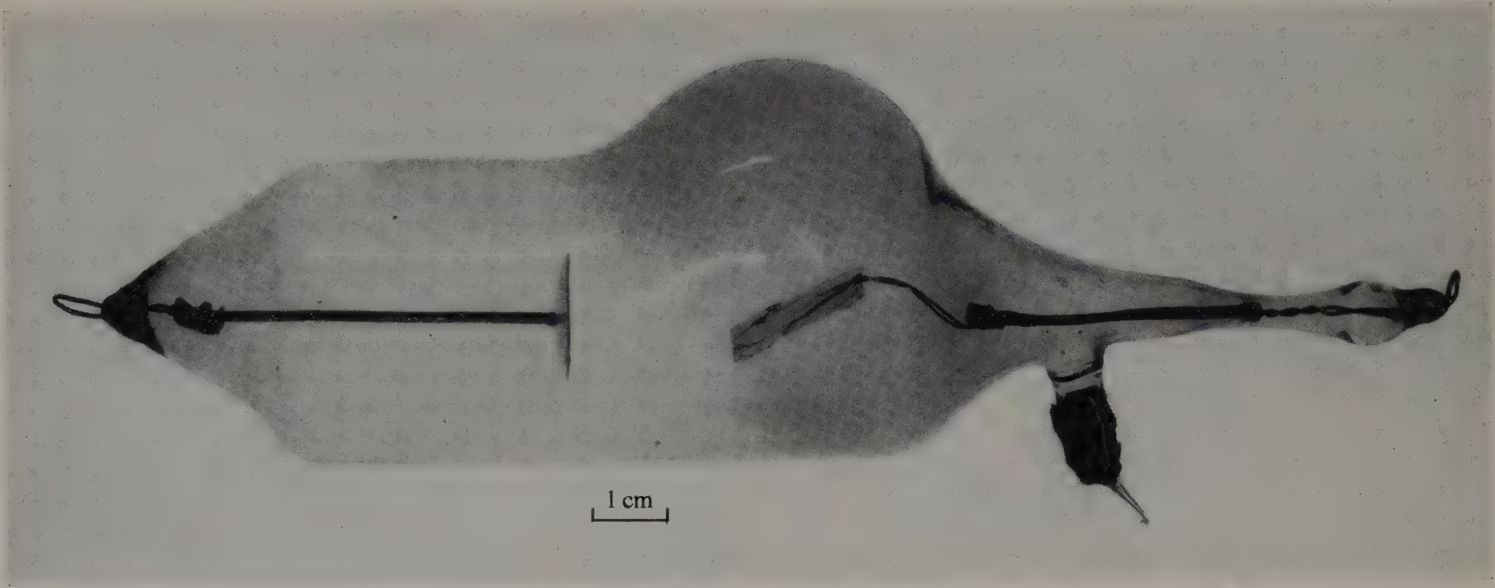
Radiation

If a solid or liquid is heated sufficiently, it will (if it does not burn or boil away) become first red-hot and then white-hot. These are signs of a more general effect, namely radiation from bodies at any temperature. Indeed, one can detect the heat from a domestic iron long before it is red-hot by putting one's hand near it. All objects are supposed to radiate, including one's own body; the loss of heat from unprotected parts of the body on a cold day is partly due to radiation. There is, however, no net transfer of heat by radiation between bodies all at the same temperature in a closed room: although some bodies in fact radiate more than others they balance this loss by absorbing more of the radiation that falls on them. The best radiator is also one that absorbs all the radiation it receives; it is the ideal 'black body', and the radiation it emits is called 'black-body' radiation. It is identical in quality with that which escapes from a small test hole in the wall of an enclosure of any material or materials at constant temperature, and this is the best way of studying it. This independence of material shows that black-body radiation is something fundamental. A good deal was known about this radiation before 1895, notably that its total intensity varies as the fourth power of the absolute (Kelvin) temperature. It was also known that the curve expressing intensity of radiation against wavelength has a maximum; and that the wavelength at which this maximum occurs varies inversely as the absolute temperature, as Wien discovered in 1893. The first fact could be explained thermodynamically, and so could the last if it was assumed that a maximum exists, but this observed fact was proved to be contrary, not indeed to thermodynamics, but to some apparently very well-founded ideas, known as the equipartition of energy, when they were applied to the electromagnetic theory of light and other radiations. The discrepancy was no minor one: there ought not to be a maximum at all; the intensity curve should rise without limit to the shorter wavelengths; and all the energy save an infinitesimal fraction should be of very short wavelength. In other words, ordinary visible black-body radiation should not exist!

In the year 1900 Max Planck published a paper that is a major landmark in the history of science. He succeeded in finding a mathematical formula that fitted the experimental facts, and then proceeded to examine what it implied physically. He came to the conclusion that it was necessary to suppose that radiation was emitted, and probably absorbed also, discontinuously in what he called 'quanta'. Radiation of any particular frequency, say ν, could be emitted only in amounts $h\nu$ and so on. The symbol h was a constant whose value he could find by comparing his theoretical curve with the experimental one (Fig. 3.6).†

It was a long time before this idea was fully accepted. Pretty clearly Planck had got the right formula, but the more conservative physicists kept hoping that some less drastic way could be found of explaining it.

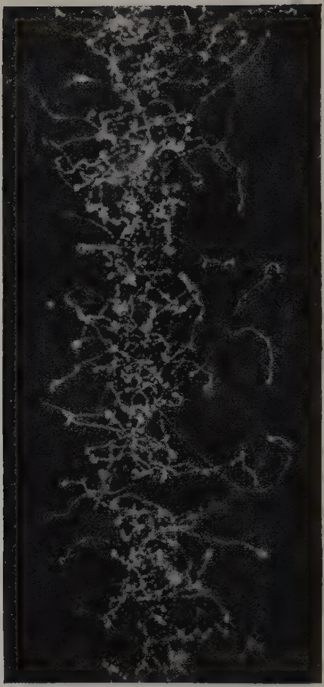
† $h = 6{\cdot}55 \times 10^{-27}$ in erg seconds. Its magnitude on the atomic scale can be better realized by saying that the energy of a quantum of green light is that acquired by an electron accelerated through about 3 volts.



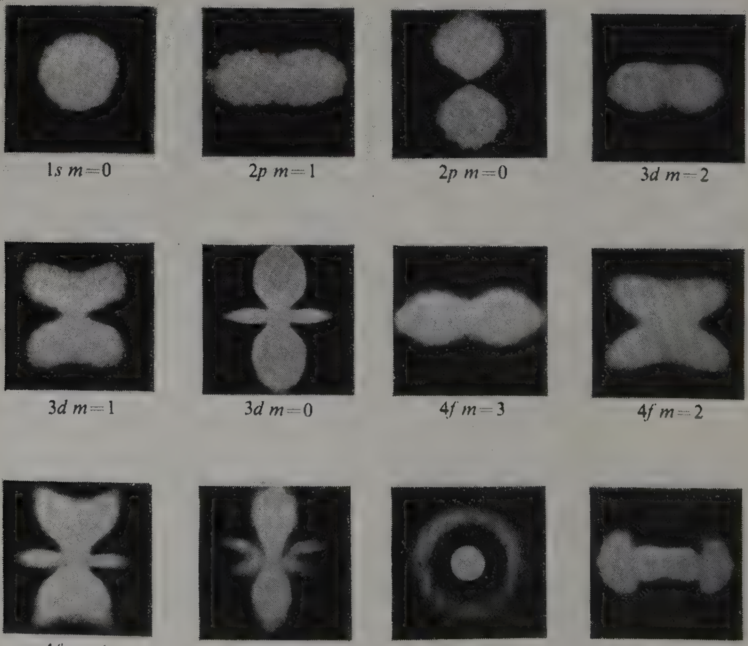
Early X-ray tube now in the museum of the Cavendish Laboratory. The cathode is the disc to the left. The inclined foil to the right is the target and source of the X-rays. It serves also as an anode. Many tubes have a separate anode. From G. P. Thomson, *J. J. Thomson and the Cavendish Laboratory* (Nelson, 1964).

2

(a) Wilson tracks due to X-rays. The X-rays were limited to a beam. Each of the tangled tracks contains many drops but is due to only one electron. From C. T. R. Wilson, *Proc. R. Soc.* A **87**, Pl. 8 (1912).



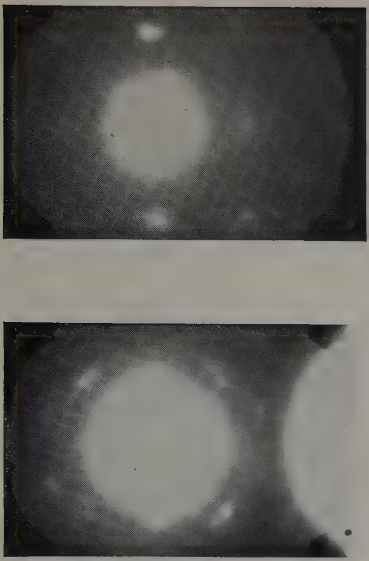
(b) Distribution of charge in a hydrogen atom. These pictures show in silhouette the intensity of the electronic wave, i.e. the charge density in a hydrogen atom excited in various ways. The top picture on the left represents the ground (normal) state. From M. Born, *Atomic Physics* (Blackie, 1944).



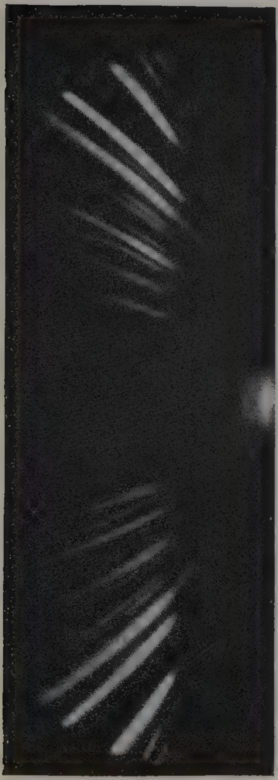
$1s\ m=0$ $2p\ m=1$ $2p\ m=0$ $3d\ m=2$

$3d\ m=1$ $3d\ m=0$ $4f\ m=3$ $4f\ m=2$

$4f\ m=1$ $4f\ m=0$ $2s\ m=0$ $3p\ m=1$

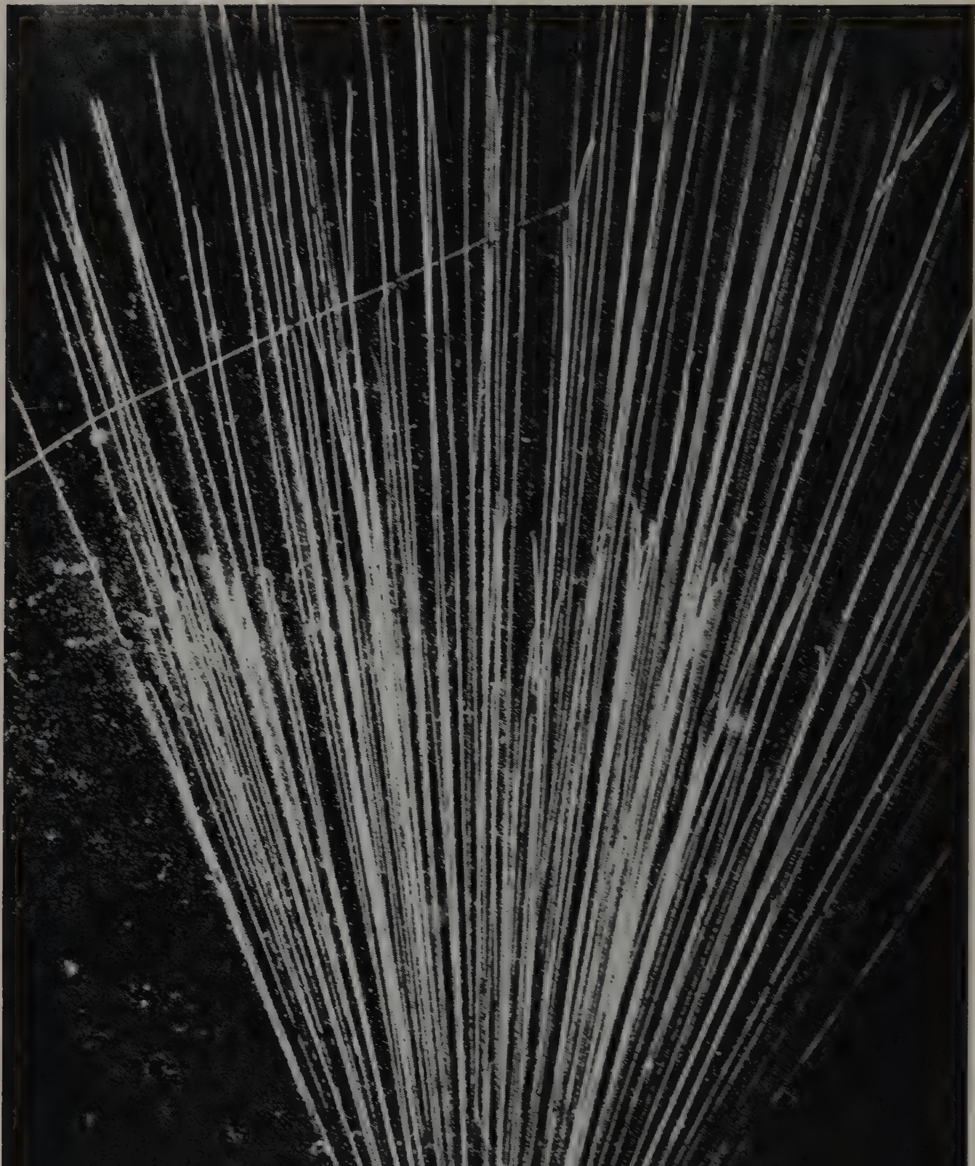


(a) Early electron diffraction pattern. This aluminium specimen had only a few crystals in the area used, one of which caused most of the spots. The aluminium atoms in the planes normal to the electron beam were in a square array.

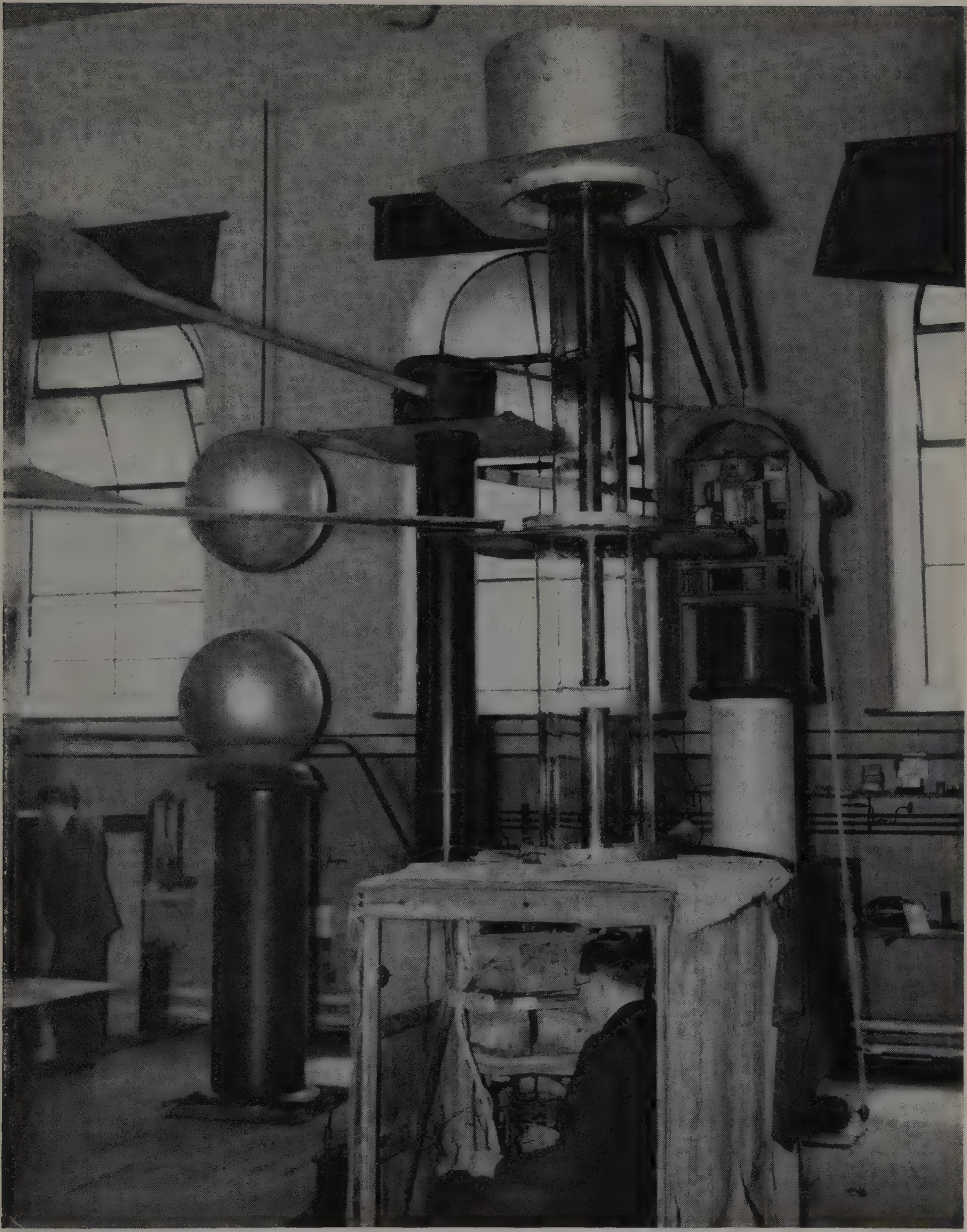


(b) Positive rays. Pattern made on a photographic plate by positive rays that have been deflected horizontally by an electric field and vertically by a magnetic one. The pattern is doubled because the magnetic field was reversed half-way through the exposure. Each parabolic arc represents one value of e/m and one kind of atom or molecule. The three brightest on each side correspond to the atoms of oxygen, nitrogen, and carbon with unit charge. Weaker areas further from the centre correspond to some of these with double charges. The gas was the residual air with an impurity of hydrocarbon from the grease used for the glass taps.

4

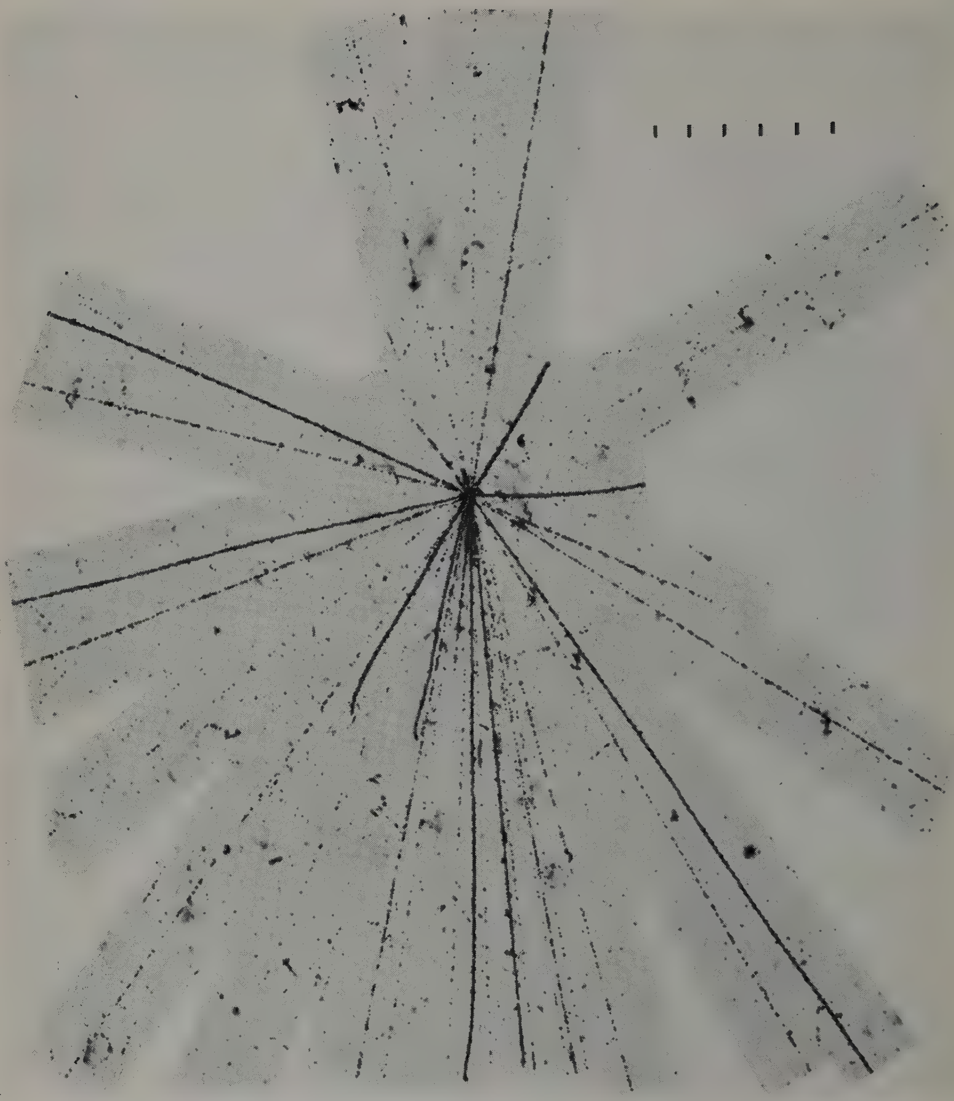


Disintegration of a nitrogen nucleus by an α-particle. A proton of very long range is emitted. α-particles of two different ranges are seen, emitted by a mixture of thorium B and C. From *Proc. R. Soc.* A **136**, P. M. S. Blackett.



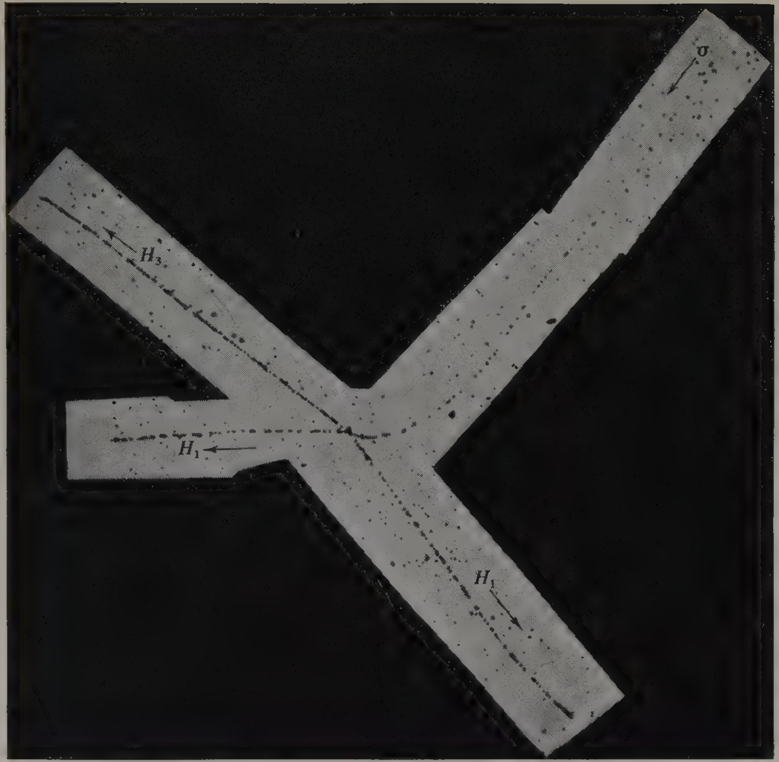
Cockroft and Walton's apparatus.
From *Les prix Nobel en* 1951.

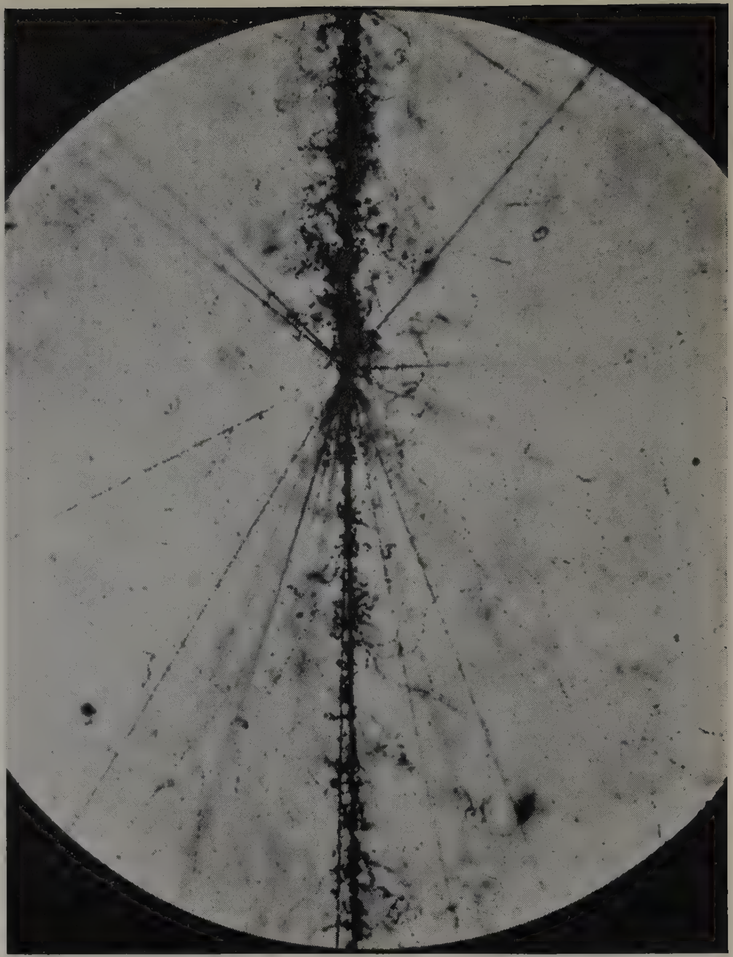
6



Disintegration of a silver or bromine nucleus in the plate by a proton of energy 3×10^{10} V. The proton enters from the top of the picture. From C. F. Powell, P. H. Fowler, and D. H. Perkins, *Study of elementary particles by the photographic method* (Pergamon Press, 1959).

(a)



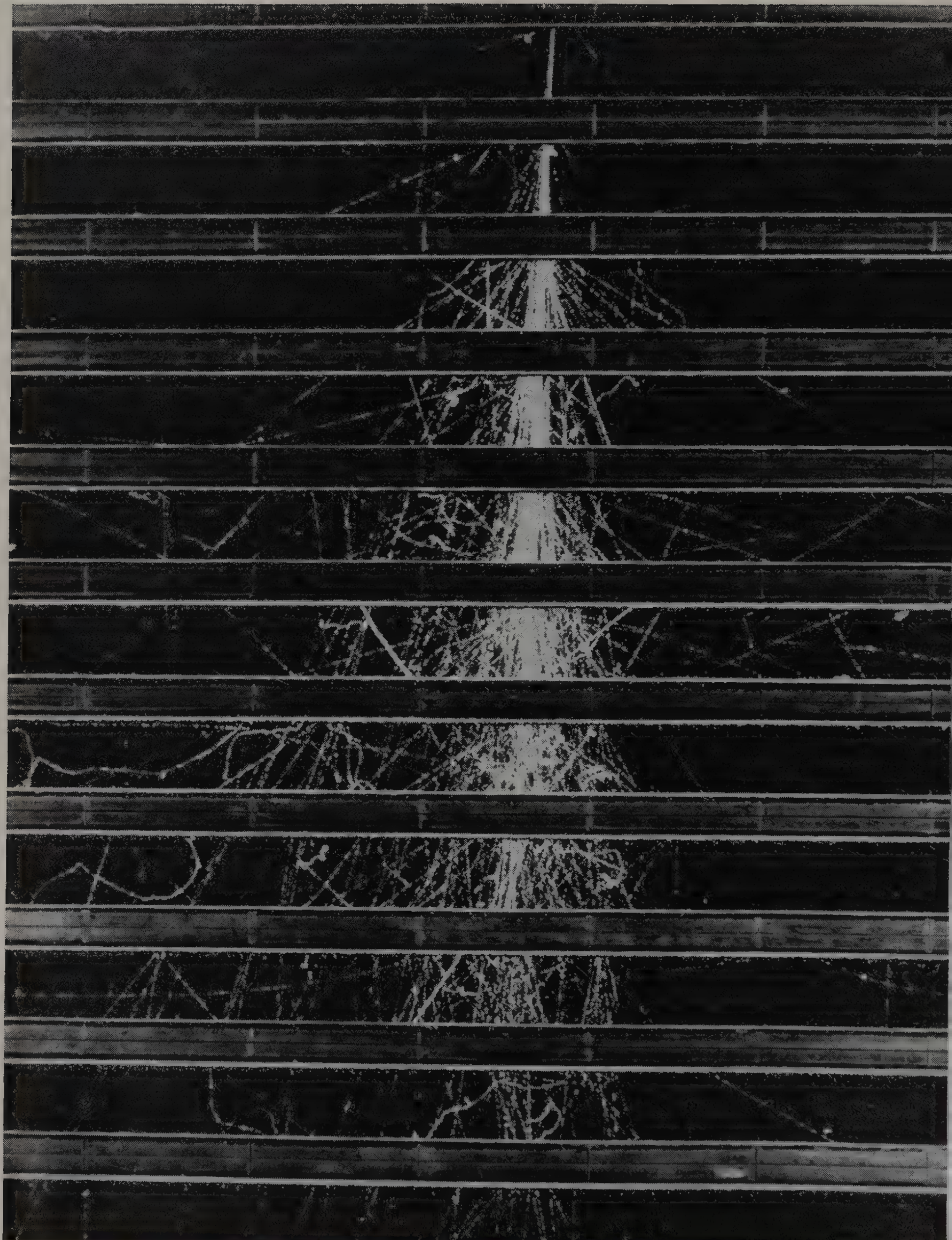


(b)

(a) First observation of a disintegration by a negative meson. The meson comes from the top right side. The track at top left is probably an H_3; the other two exit tracks are H. From C. F. Powell, P. H. Fowler, and D. H. Perkins, *Study of elementary particles by the photographic method* (Pergamon Press, 1959).

(b) Collision of a nucleus of iron. The cosmic ray comes from above; note the very dense track it causes. The roughnesses are due to secondary particles. From C. F. Powell, P. H. Fowler, and D. H. Perkins, *Study of elementary particles by the photographic method* (Pergamon Press, 1959).

8



Cosmic ray cascade. A shower developing through a number of brass plates 1·25-cm thick placed across a cloud chamber. The shower was initiated in the top plate by an incident high-energy electron or photon. Photograph by the M.I.T. cosmic-ray group. From Rossi, *Cosmic rays* (McGraw-Hill, 1966).

Belief came when it was found that the quantum theory could account for other difficulties not obviously connected with black-body radiation. The first of these concerned specific heats of solids. It had long been

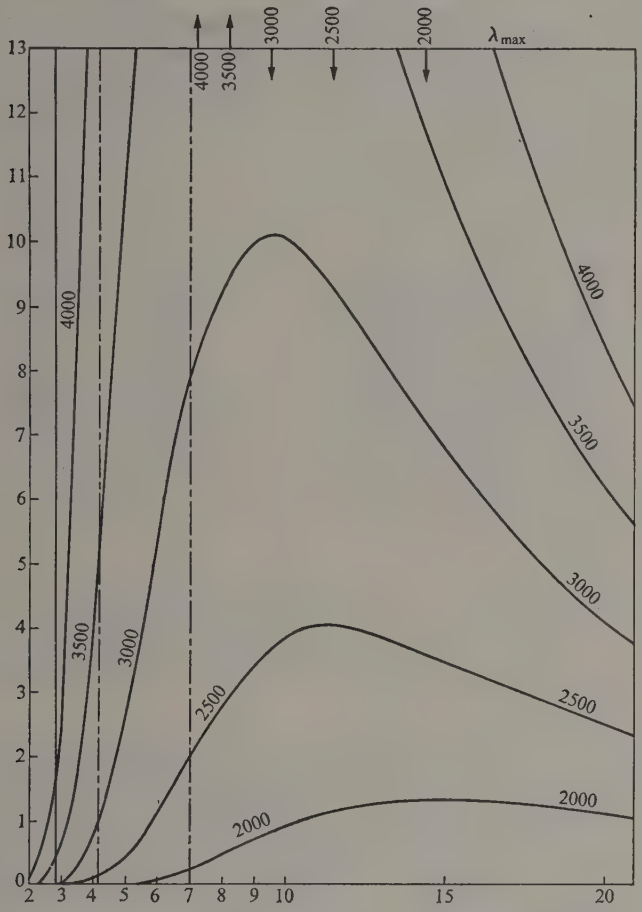


FIG. 3.6. Intensity of black-body radiation on Planck's formula. Intensity as function of wavelength is 10^{-5}cm. The visible region is between the dotted lines. Note the wavelengths for maximum intensity at the top of the diagram. From M. Born, *Atomic physics* (Blackie, 1944).

known that the amount of heat required to raise the temperature of elements in the solid state by one degree was nearly the same, reckoned not per gramme but per atom. This agreed nicely with the kinetic theory principle of the equipartition of energy. One example of this very general principle is that in a mixture of two gases, both monatomic, the average kinetic energy of the atoms of one gas is equal to that of the other. Maxwell proved this on Newtonian mechanics and it seemed highly probable that it was also true for the atoms of solids in thermal equilibrium. However, in solids the atoms vibrate about fixed positions because of their heat energy; and it can be proved, again on Newtonian mechanics, that as in the case of a pendulum, or of a weight oscillating up and down under the force of a spring, the average kinetic energy is equal to the average potential energy of the forces that hold the atom in place. Thus the heat energy that has to be added to cause a one-degree rise of temperature is twice what it would be if the atoms were free as in a gas. For the heavier elements this gave roughly the observed results, but for some of the lighter ones the heat required was substantially less than it ought to be. Carbon in the form of diamond was an especially bad case, and the trouble was worse at low temperatures.

Einstein showed that this could be explained by supposing that the energy of each vibrating atom is quantized and can occur only in amounts $E=h\nu$ times a whole number, where ν is now the frequency of the vibration and h is the constant of the quantum theory. Planck had found a mathematical expression for the way the average energy of a quantized oscillator varies with temperature, and applying this Einstein deduced an expression for the specific heat at any temperature in terms of one frequency ν_c, characteristic of the particular substance. This worked fairly well. Later refinements taking account of the joint motion of groups of atoms have improved the fit, and this problem seems to be solved in terms of Planck's theory.

Another application of Planck's idea, this time in a modified form, was also due to Einstein (1905). In the photoelectric effect (p. 62) light causes the emission of electrons from certain metals. The electrons have varying energies but Lenard had earlier proved that the maximum energy of projection is independent of the intensity of the light though the number of electrons is proportional to it. The maximum energy increases with the frequency of the light, i.e. as one moves through the spectrum from red to violet and beyond. The electrons appear immediately the light reaches the metal plate; so there can be no accumulation of energy in the plate. The odd thing is the production of electrons of considerable energy, though few in number, by very faint light, the limit being set merely by the sensitivity of the detector. In fact, electrons appear when the intensity is so low and the time so short that the energy of the electrons must be gathered from the radiation striking an area many times that of an atom, *assuming the energy to be evenly distributed over the irradiated* *surface.* It is difficult to see how the electron can gather in energy from so large an area. Einstein denied the assumption and instead supposed the energy of the radiation to be concentrated into quanta, not only in emission and absorption but also in the region between. These moving quanta are now called 'photons'. Their energy for monochromatic light is assumed to be *hν only*; 2*hν*, 3*hν*, etc. do *not* appear. Millikan later showed that Einstein's assumption fully agrees with the facts for photoelectric emission. It leads, however, at once to a terrible dilemma.

A particle theory of light, not so unlike Einstein's, had been put forward by Newton in the seventeenth century but completely rejected in favour of the rival wave theory as a consequence of the work of Young and Fresnel early in the nineteenth century. There is a wide range of experiments in optics dealing with interference and diffraction, as they are called, capable of high accuracy and often of considerable beauty. The explanations of these experiments depend fundamentally on assuming that two beams of light ultimately derived from the same source and brought back together can wholly or partially destroy one another's effect in certain places while reinforcing it in others. This is credible for waves, since the crest of one may neutralize the trough of another and vice versa (Fig. 3.7). It is incredible that one set of bullets can neutralize the effect of another set going at only a small angle to it, as these experiments often imply. It is also necessary to assume that light passing close to a material object is slightly deflected by it. This is to be expected for waves; waves of sound, for example, can go round corners, though with

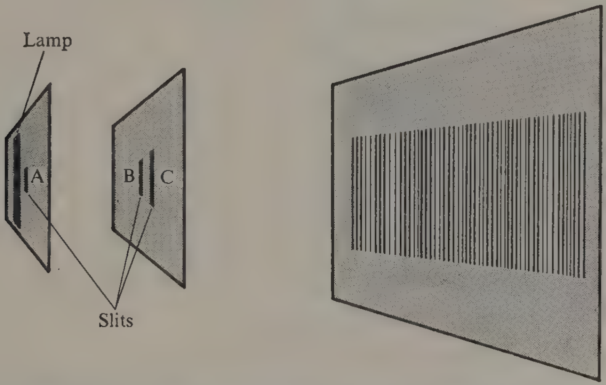


FIG. 3.7. Young's interference experiment. Slit A is needed to get a fine beam of light, which then goes through slits B and C, also very fine. The beams AB, AC spread horizontally by diffraction and overlap on the last screen giving the bands as shown (the width and spacing of these are exaggerated for clearness).

diminished intensity. It could at first sight be explained for particles, and was so explained by Newton, by supposing that matter exerts a force on light passing very near it, but this explanation did not fit the details of the experiments. The wave theory seemed one of the best established in physics, though the nature of the medium carrying the waves was still a matter for discussion.

This difficulty got worse and worse the more physics advanced. X-rays were found to present the same difficulties as light but in an even more violent degree. The ionization caused by X-rays was demonstrated in various ways, but most conclusively by C. T. R. Wilson in his cloud chamber (Fig. 3.8 and Pl. 2(a)), to be mostly indirect. The X-rays produce high-velocity electrons from a few atoms in the gas or the walls of the chamber, and these ionize a great many more by colliding with them, behaving very much like cathode rays. This is an exaggerated sort of photoelectric effect. But the Braggs had proved that X-rays had wavelengths measurable by using crystals in place of mechanically ruled diffraction gratings and had used the diffraction patterns of these crystals

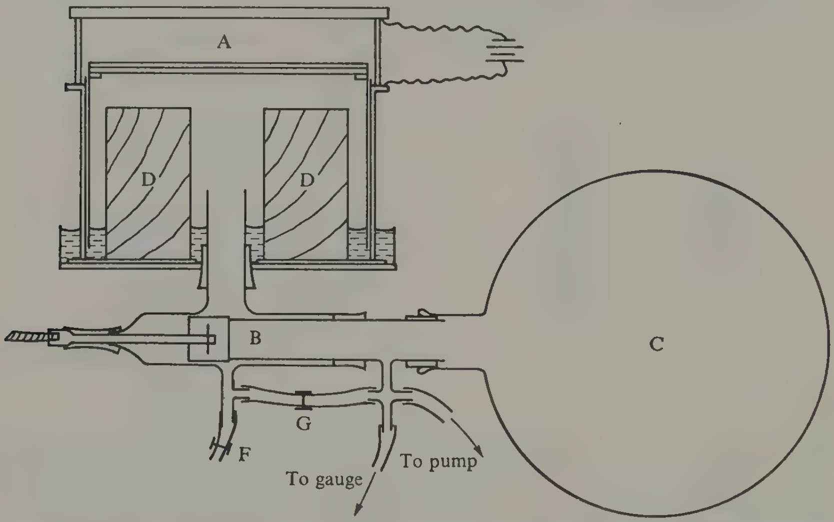


FIG. 3.8. Wilson cloud chamber. *A* is the actual chamber. The expansion is effected by opening the valve *B* and so putting the air-space below the plunger in communication with the vaccum chamber *C*. The floor of the chamber drops suddenly until the skirt of the plunger strikes the indiarubber-covered base-plate. The wooden cylinder *D*, by taking up space, reduces the amount of air that has to go through the tubes at the expansion. When the moist air in the chamber expands it cools and tries to condense, but in the absence of dust can only do so on to charged ions, each of which becomes the centre of a drop when the conditions are well chosen. From C. T. R. Wilson (1912) *Proc. R. Soc.* A **87**, 278.

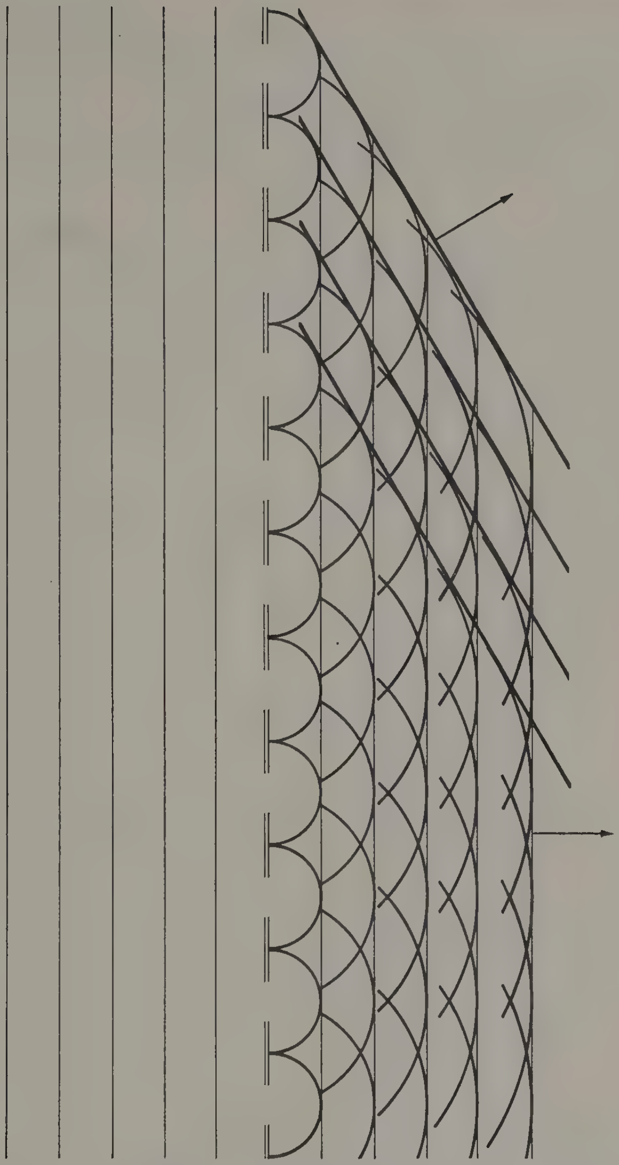


FIG. 3.9. Principle of a diffraction grating. A plane wave whose crests are shown in fine lines advances from the left to a screen pierced by slits perpendicular to the paper and at regular intervals. Wavelets start from each slit, all in phase. Parts of the crests of some of these wavelets are drawn. They form nearly straight lines parallel to the original crest which represent a wave going straight through and also one going obliquely upwards, which is a *diffracted* wave. (Another going obliquely downward is omitted to avoid confusion.)

to find the arrangement of the atoms in them, all in complete accordance with wave theory (Fig. 3.9). This mystery, not unnaturally, caused an increasing tension in physicists and the history of physics from about 1905 to 1925 or later must be read with this in mind.

One other line of work must be mentioned briefly before we turn to the direct influence of quantum theory on the structures thought probable for atoms.

From quite early days (*c.* 1900) it had been known, largely because of Townsend's work, that fast-moving electrons could detach electrons from atoms with which they collided, and that this process played an important part in the very complex phenomena of sparks and other electric discharges. After losing an electron an atom will have a positive charge that will attract the electron back, and it is clear that the electron will escape only if a certain minimum of energy is available. When the ionization is produced by the impact of an electron, this electron must, to begin with, have enough energy to carry both itself and an electron originally part of the atom clear of the attraction of the atom. Various attempts had been made to measure this energy, usually expressed as the voltage, called the 'ionization potential', that must accelerate an electron in order that the electron may gain it. However, though the results gave the right order of magnitude they were very inaccurate, partly because another effect confuses the experiments. This was discovered by Frank and Hertz (1914) in mercury vapour. Here, as in the inert gases, an electron remains free in spite of collisions, unlike what occurs in most of the commoner gases where sooner or later it will stick to an atom or molecule to form a heavy negative ion. Frank and Hertz found that the electronic collisions were elastic until the electron had a certain critical energy, when collisions became possible in which it lost the whole. Electrons with still greater initial energy lost the critical energy and retained the difference. They also found that at and above this critical energy light appeared in the vapour, whose spectrum was a single line of the mercury spectrum far in the ultra-violet. Furthermore, the frequency ν of this line is such that $h\nu$ is equal to the critical energy of the electron. This is a kind of reversal of Einstein's view of the photoelectric effect: there, a photon gives energy to an atom producing an electron; here, an electron gives energy to an atom that radiates it as, presumably, a photon. A little later it was proved that the mercury atom is *not* ionized by electrons of this energy (4·9 V); in fact, 10·4 V are required. A few months before these experiments this possibility had been suggested by Niels Bohr in a paper that introduced a new era in physics.

Atomic structure

The problem of the structure of atoms, and especially of the arrangement of the electrons inside them, had occupied the thoughts of physicists ever since the discovery of the electron. Indeed, in his first paper J. J. Thomson suggested concentric rings or layers of electrons, the only evidence being that afforded by the periodic table of the elements, which suggested that as the weight of the atom, and so presumably the number of electrons, increases the chemical properties change cyclically. This makes sense if the latter depend mostly on the electrons in an outer layer whose numbers should vary from 0 to 8, the extra electrons going inside, perhaps in more than one layer. In fact, as we shall see, this chemical evidence pointed to the truth, but it took physics thirty years to get there.

The reader may perhaps wonder why it took as long as it did for the idea of an atom as a miniature planetary system to be adopted. There were in fact good reasons against it. First, atoms have a fairly definite size, as the difficulty of compressing solids and even liquids shows. If not as hard as cannon-balls they behave more like tennis balls than, say, grapes. What determines this size? As Hertz and Lenard showed, electrons can go through solids, i.e. through atoms without tearing holes. They can hardly be supposed to pack together in an atom as the atoms do in a solid. But if the only forces holding the atom together are electrical there is nothing to fix the size. All one has is the charge on the electron, its mass, and the velocity of light; and these by themselves cannot fix a length. The charge on the lightest nucleus, that of hydrogen, which was by 1913 beginning to be called the proton, was of course the same with sign changed; the mass of the proton merely gives a ratio to that of the electron of 1836 or so but not a length. Any atomic arrangement could, in fact, be doubled or halved in size and work just as well. It is the same with gravity. The scale of the system of planets moving round the sun was fixed by some series of events in the remote past; it is, so to speak, a matter of geography, not of physics.

The other difficulty was that of radiation. According to Maxwell, any charged body, except one at rest or moving with uniform velocity in a straight line, will radiate energy; light, in fact, using the word to include the whole spectrum from long-range radio to gamma rays. As the moving body loses energy it will spiral in towards the ‘sun’ and the atom will collapse. If there were a ring of electrons following one another closely in the same orbit, radiation would indeed be enormously reduced, though hardly enough to allow atoms to last for millions of years. Maxwell, in an essay on Saturn's rings, had shown that while these are stable it is only because the small objects of which they are composed attract one another. If they repelled, as of course electrons would, the rings would be unstable. Instability also kills the idea of a static atom. If the forces between particles are purely electrostatic, every arrangement can be proved unstable.

Thomson got round the first difficulty by postulating that the positive electric charge formed a spherical jelly whose size fixed that of the atom. On certain assumptions this gives stability also, but the idea was artificial and was completely killed by Rutherford's discovery of the nucleus. This discovery, however, made the above difficulties all the more glaring. If they had not been acute it is unlikely that Niels Bohr's theory would have received even the doubtful support it did when it appeared in 1913. Bohr was then working in Rutherford's laboratory at Manchester as a theorist, though he had done some experimental work in Denmark.

His theory considered the simplest atom, that of hydrogen, assumed to consist of an electron moving in a circle round a proton. Bohr begins his paper by calling attention to the difficulty of fixing the size of an atom, and saying that this can be done if in fact it depends on Planck's h. He then considers the emission of light from a hydrogen atom, considering first the light that would be emitted, assumed to be a single photon, when an electron moves from a great distance to end up in the stable state of the atom. He assumes that any stable or semi-permanent state of the atom must satisfy ordinary mechanical and electromagnetic principles but that the transitions which produce the photons do not. He thus openly denies the validity of the instability due to radiation, which for a single electron is the only one that arises. It was known that the spectrum of the hydrogen atom was one of sharp lines, i.e. that the light consisted of certain definite frequencies and that these frequencies were given by two mathematical expressions, called the Balmer and Paschen formulae, one for the visible and one for the infra-red parts of the spectrum. Since the frequencies were sharp the energy of the photons emitted must be definite, for Bohr assumed that $E = h\nu$ is to hold. Since there were many lines there must be many energies. Both the Balmer and Paschen formulae express the frequencies as *differences* of two sets of terms. If these terms were 'energy levels' (such as those that Frank and Hertz were in process of discovering while Bohr was writing) these formulae could be explained, since transitions between the levels in each series would give frequencies equal to the difference in energy of pairs of levels divided by h. It remained to calculate the 'energy levels'. To do so a third assumption was needed. In making this Bohr was guided by the principle that the results of the quantum theory must approach those of the older physics when many quanta are involved. The assumption he made was in effect that the allowed energies correspond to circular orbits for which the product of momentum and radius is a multiple† of $h/2\pi$ (this was quickly extended and modified to include elliptical orbits like those of the planets). With these circular orbits the radius goes as the square of the 'quantum number', i.e. the multiple of $h/2\pi$ considered. Thus the orbits are concentric circles, and if the radius of the inside one is taken as unity that of the second is 4, that of the third 9, that of the fourth 16, and so on. The larger radii thus greatly exceed the ordinary size of the atom, and spectral lines corresponding to transitions from them are in fact not seen unless the density of the luminous gas is very low. Electrons in these orbits become entangled with the neighbouring atoms unless these are very far apart. The radiation given by a transition between states whose quantum numbers are large and differ by unity has according to this theory a frequency equal to the frequencies, which are then nearly the same, with which the electron describes the orbits. This fits the original idea that for large quantum numbers the quantum theory gives results that approximate to those of established electrodynamics, since the frequency of the light emitted from an electron moving round in a circle would on the established theory be the actual frequency of the revolution.

Bohr called this idea the 'correspondence principle' and it has remained valid throughout the later changes in quantum theory.

The assumptions of Bohr's theory were startling, but it not only accounted exactly for the spectrum of hydrogen in the visible and infrared, both in the nature of the series and in the absolute values of the frequencies, but predicted another series of lines that Lyman shortly after found in the ultra-violet. Still better, Bohr was able to show that certain spectra then also ascribed to hydrogen could be better assigned to helium if that atom had lost one of its two electrons and had the other displaced. This lone electron would then behave like the lone hydrogen electron except that it was subject to twice the force, since the helium nucleus has a double charge. This is not quite true, for there should be a

† The presence of the 2π term is *not* a mysterious fact of nature but a matter of definition. If Bohr had taken the product of momentum and circumference it would not have come in; he chose as he did to fit in with existing definitions in Newtonian mechanics. In Bohr's paper another mathematical assumption is originally taken which is shown later on in the same paper to be equivalent to that stated. The whole paper is an admirable example of the fact that discoveries are usually made the hardest way round. It is much easier if you know the answer!

minute difference, predicted by Bohr and confirmed by experiment, which comes in because neither the proton nor the helium nucleus is strictly infinite in mass compared with the electron and both wobble slightly to different degrees as the electron revolves.

The size predicted for the smallest orbit, from which by postulate no further radiation is possible, was about that expected for the atom. Acceptance of Bohr's theory was hindered by the outbreak of war but in spite of the boldness, not so say inconsistency, of its assumptions it made rapid progress after peace. The years from 1919 to 1925 were spent by a large proportion of physicists led by Bohr, Sommerfeld, and Born in generalizing the theory and applying it to the changes in spectra produced by electric and magnetic fields (the Stark and Zeeman effects) and extending it to other substances. Moseley, killed at Gallipoli, had already shown that when applied to the inner electrons of heavy atoms it accounted for the X-ray spectra just measured by Sir William Bragg. The successes were great, but there were difficulties in detail and the conceptual difficulties of the theory were not really reduced, though time softened their apparent crudeness. Towards the end of this time it was found necessary to suppose that the electron is more complicated than a tiny sphere, that it behaves in some ways like a top, but a rather queer top. It was called the 'spinning electron'.

Electrons in atoms

By 1924 the position as regards the structure of atoms was thus only moderately satisfactory. Qualitative, and in certain cases even quantitative, explanations had been given for many of the simpler atomic spectra and for the ways in which the lines were split up in magnetic and electric fields. It had also been shown by Stern and Gerlach that atoms moving in a good vacuum could be orientated by a magnetic field and when so orientated were pulled sideways by the field as small moving magnets would be: the atoms in fact became magnets with polarities directed at certain definite angles to the lines of force of the field, in the simplest case either with it or against it. This was called 'space quantization', and confirmed the quantum theory of the behaviour of atoms in a magnetic field as deduced from their spectra.

However, in spite of it all there were difficulties. The frequency of the periodic motion of the electrons in their orbits remained mysteriously unobservable except in the limiting case of large quantum numbers, when it agreed with Bohr's correspondence principle (see p. 65). The arrangement of the electrons in the more complicated atoms, in so far as it was known, seemed decidedly arbitrary.

The second difficulty was solved first, at least formally, by a new principle stated by Pauli in 1925. Pauli assumed that the state of every electron was determined by a set of four integers (or half-integers) called *quantum numbers* and asserted that only one electron could be alloted to any given set; hence the name 'Pauli exclusion principle'. In Bohr's original paper there had only been one such quantum number determining the radius of the orbit, which was assumed to be a circle. A point requires three coordinates to determine its position in space, and the orbit of a point electron moving in a known field of force (e.g. from the nucleus) would require the same number to fix it. The quantum theory came in by asserting that, when suitably reckoned, these coordinates had to be integers. But, as has been said, the development of Bohr's theory had necessitated the introduction of the 'spinning electron'. Three quantum numbers were not enough. Pauli maintained that a fourth quantum number was needed but that it should be allowed only two values, $\pm\frac{1}{2}$. The use of $\frac{1}{2}$ rather than unity simplifies some formulae; the essential is that it can have only two values.

One quantum number, called the *principal quantum number*, is the largest, and for those electrons that lie near the nucleus it determines the mean position and energy. The electrons near the nucleus are more tightly bound and normally have a lower principal quantum number than those farther out. But among the outer electrons of the heavier atoms it is possible for an electron of lower principal quantum number to have a higher energy, be less tightly bound, and spend most of its time farther from the nucleus, than one of higher quantum number.

On Pauli's principle only one electron can exist in the atom in the state defined by a particular set of four quantum numbers. The last quantum number, called the *spin quantum number*, usually has rather little effect on the energy, and it is convenient to alter the principle to say that at most *two* electrons can exist in a state defined by the first *three* quantum numbers. There are certain rules governing the possible values that the second and third quantum numbers can take; these would take too long to detail, but they lead to the rule that the greatest number of electrons allowed with principal quantum number n is $2n^2$. Thus for $n=1$ two atoms are possible, hydrogen with 1 electron and helium with 2. With $n=2$ the formula gives 8 and covers the elements from lithium to neon inclusive. Neon, which closes the set of 8, is an inert gas like helium, which closes the set with $n = 1$. Since $2 \times 3^2 = 18$, this is the number of kinds of atoms in the third set and the eighteenth element after neon is krypton, also an inert gas. However, the simple regularity breaks down after argon, the eighteenth element in the whole list, because some electrons with principal quantum number 4 are preferred over those with quantum number 3 for the reasons given above. Though this scheme of Pauli's was stated originally for the orbit theory of atomic structure it has survived the change to the wave theory, with which we must now deal, and it therefore seems appropriate to give it here.

Wave mechanics

Pauli's principle put inorganic chemistry on a reasonable foundation, though a good deal still had to be taken on trust, but the difficulty of the unobservable orbits remained, as well as certain minor discrepancies between theory and experiment. Worst of all, however, was the paradoxical behaviour of light, now as waves, now as particles. All these were overcome in the years from 1924 to 1927, but at the cost of profoundly altering the philosophical basis of physics from the prediction of what *must* happen in particular circumstances to that of what is *likely* to happen.

Two lines of attack, apparently very different, one led by Heisenberg, the other by de Broglie, appeared almost simultaneously in 1924. In fact, notwithstanding their different approach, they led to the same conclusions and have now been amalgamated. As the de Broglie approach is more easily expressed in non-mathematical terms than the Heisenberg one, I shall take it in what follows.

De Broglie attempted to resolve the conflict between waves and quanta in radiation by supposing that radiation consists of particles of vanishingly small 'rest mass' (see 'relativity', Chapter 2) moving with a speed differing only infinitesimally from that of light. This resulted, using the theory of relativity, in their having a finite energy, which de Broglie took as $h\nu$, ν being the frequency of a vibration inherent in the particle *as observed from the laboratory*. This can provide the photons that the experiments on photoelectric effects and ionization by X-rays require, if we suppose that ν is also the frequency of the light as calculated from the velocity of light and its wavelength as measured by the experiments on interference and diffraction. De Broglie claimed that such an inherent vibration in a particle moving with a speed near that of light would on the theory of relativity appear to a laboratory observer as a wave, which could be identified with the wave as observed and measured in the interference experiments if it were supposed that the wave in some way controlled the particle so that the particle appeared only when the wave was strong. De Broglie never in fact fully explained how this guiding was to be done or even that if it were the results would agree with all the experiments.

However, there was more to the theory than has been said. De Broglie claimed from his earliest English paper (an extract from his doctoral thesis at the University of Paris published in the *Philosophical Magazine* in 1924) that this theory applied to *all* particles, in particular to electrons, and that it explained the otherwise arbitrary choice of the permitted orbits of Bohr's theory. In fact, these orbits contain an exact number of wavelengths of the electrons moving round in them (Fig. 3.10). Thus, the wave going round one of these orbits returns on itself and joins up smoothly, rather as the pattern of a wallpaper will join up correctly round a room if the total length of the walls is exactly a whole number of units of pattern.

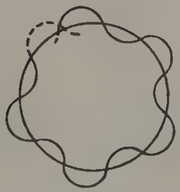


FIG. 3.10. When the wavelength is right the waves fit round the circle. A single definite wave form its possible only when the circumference of the circle is a whole number of times the wavelength. From M. Born, *Atomic physics* (Blackie, 1944).

De Broglie's theory was a brilliant sketch rather than a finished plan. Schrödinger expressed it, or at least that part of it relating to atomic structure, in a form from which precise deductions could be made and applied it to the hydrogen atom. It agreed with the orbit theory in almost all respects except those needing the spin quantum number. This exception was not surprising, for Schrödinger's theory was avowedly an approximation neglecting relativity corrections. Its merit was that the unobservable orbits had disappeared with some of the minor differences and the electron was represented as a kind of smudge, or rather a set of smudges each roughly representing previous orbits. Schrödinger was inclined to regard these 'smudges' as continuous distributions of the electron's charge and mass, but this idea broke down when applied to free electrons; for one thing, the 'smudges' in this case would spread indefinitely with time. Born replaced it with the idea that the intensities measured the *probability* of the electron being found in particular regions of space (Pl. 2(b)).

Electron diffraction

This way of looking at things applies particularly well to the experiments of Davisson and Germer and of the author and Reid on the behaviour of electrons reflected from, or passing through, crystals. The atoms of a crystal are arranged in regular order and Laue had shown in 1913 that they behave for X-rays in a way similar to that in which a finely ruled grating of glass or metal—a diffraction grating—does for ordinary light, except that the array of atoms in a crystal has a three-dimensional

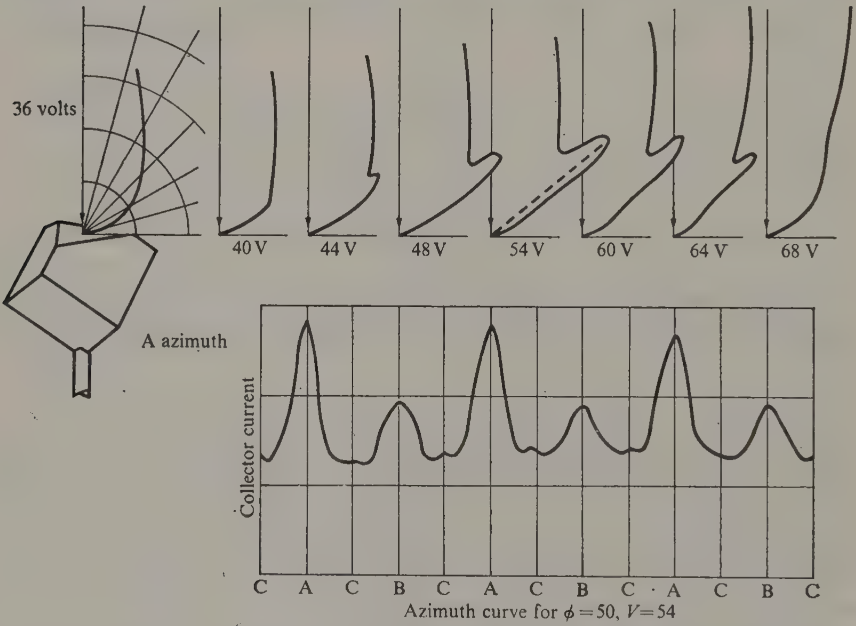


FIG. 3.11. Electron diffraction from nickel crystal. Electrons are scattered from the octahedral face of a single crystal of nickel. When the voltage is adjusted, in this case to 54V, and the angle of the collector to 50° there is a strong peak of scattered electrons which repeats with the symmetry of the crystal when this is turned about the normal to the exposed plane. The relations between voltage and angle fits de Broglie's law for the grating formed by the atoms in the nickel surface. Many such peaks were found. From Davisson and Germer, in *Wave mechanics of free electrons*, ed. G. P. Thomson (McGraw-Hill, 1930).

regularity while the ruled grating repeats itself in one dimension only, namely perpendicular to the rulings. This fact, while it considerably complicates the calculations, does not alter the principles involved, and a crystal can be used to form the spectrum of a beam of X-rays as a diffraction grating does for visible light (Pl. 3(a)).

According to de Broglie's theory the wavelength associated with a particle of momentum p is h/p. For electrons of 150-V energy this is 10^{-8}cm, about the radius of an atom. Electrons with about this energy were used by Davisson and Germer; electrons of about 100 times the energy and so one-tenth the wavelength by Thomson and Reid (see Figs. 3.11, 3.12). The close agreement of these electron diffraction experiments with those for X-rays and for visible optics showed the generality of de Broglie's ideas. On Born's interpretation this means that the chance of a photon appearing in any small interval of time in any particular small region of an interference or diffraction pattern is proportional to the intensity of the radiation in that region at the time considered, which in free space is on Maxwell's theory proportional to the square of the amplitude of the pulsating electric field. In the same way the chance of an electron appearing at a given place when electrons are reflected from a crystal is proportional to the value at that place of the square of the

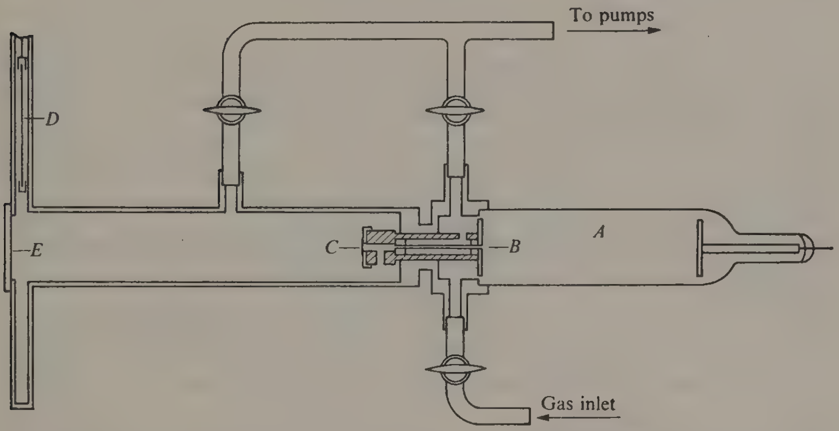


FIG. 3.12. Apparatus for electron diffraction through thin fibres. The electrons were cathode rays generated in A. A fine beam passed through the fine tube B and the specimen C of metal leaf thinned in acid or alkali, effective thickness about 2×10^{-6}cm. E was a fluorescent screen on which the diffraction patterns appeared. By lowering the plate D photographs were taken. From Davisson and Germer, *Wave mechanics of free electrons*, ed. G. P. Thomson (McGraw-Hill, 1930).

wave amplitude ψ in Schrödinger's equation. For equal wavelengths and the same crystal the patterns for X-rays and electrons would be the same but for the very important fact that the electron, having a charge, which a photon has not, interacts more strongly with matter and so is strongly absorbed. It therefore can be absorbed or scattered by a very thin layer of matter, unlike the more penetrating X-rays.

Expressed in a very crude fashion there is in each case a set of waves appropriate to any given problem; and the intensity of these waves at various places determines how many of the original lot of photons, or electrons, provided by the source would be found there.

Complementarity

An even more fundamental way of looking at these problems was called 'complementarity' by Bohr. It is concerned with the role of the observer who, after all, is an essential part of every experiment or observation. The human observer may, it is true, make use of a permanent record such as a photograph or the trace made by a moving stylus; if so, these may be regarded as taking his place. Heisenberg in 1927 introduced the 'principle of indeterminacy' deduced from his mathematical theory but illustrated by imaginary experiments. The idea is that no observation can be made without altering the system observed. Even light exerts a force on the bodies it illuminates. The pressure of light was measured by Lebedev in 1901. The most delicate probes are light—in some form—or electrons, but on wave mechanics as expounded either by Heisenberg or by Schrödinger the 'action' of a process cannot be less than about $h/2\pi$. The 'action' is a mathematically defined quantity that roughly speaking measures the change that the process implies. Thus the interference caused by an observation cannot be reduced indefinitely by reducing the intensity of the light used in the observation. This merely reduces the *number* of photons per second and so increases the time the observation will take, without reducing the shock on the system that must come from the scattering of at least the one photon that alone makes any observation possible.

What is worse, not merely is the object disturbed by observing it, but it is impossible to say what the disturbance has been. You cannot allow for the disturbance and say what the state has become after the observation. To illustrate this, take only one of the various cases considered by Heisenberg. Suppose we want to observe as accurately as possible the position on a solid of a tiny speck. It has been known for a century that to fix the position of a small object seen through a microscope requires 'light' of a small wavelength and a wide-angle lens. At the best the uncertainty cannot be less than about half a wavelength of the light used, and any limitation on the angle of the lens makes the uncertainty worse.

The pressure of light is a measure of the momentum of the photons, and one cannot use less than one photon, which must be scattered from the object into the lens if the object is to be seen. Suppose we illuminate from below (Fig. 3.13). The scattered photon will give the object a sideways impulse or recoil, which may be to either side since we cannot tell through which part of the lens it has gone. To agree with the measured pressure of light the total momentum of the photon is $h\nu$ divided by the velocity of light. The recoil for a scattering of near 90° is nearly this amount, which will be taken up by the object (or the support to which it is fixed) in the form of sideways momentum; so by measuring the sideways position, to half a wavelength one imposes a sideways momentum either way equal

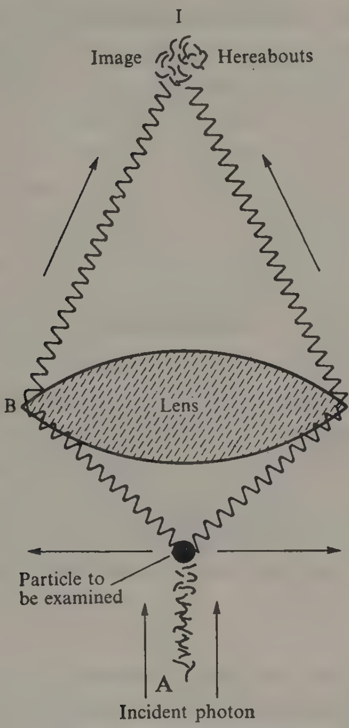


FIG. 3.13. Photon microscope. *ABI*, *ACI*, are possible paths for a photon. *ACI* will kick the electron to the left; *ABI* will kick it to the right. Note that by a well-known result in microscopy the angle *BAC* must be large if the image region is to be as small as possible, i.e. about half a wave across.

to $h\nu/c$,which is h divided by the wavelength λ. The product after the observation of the two uncertainties (in position $\Delta x = \lambda/2$ and in momentum Δp) is $\Delta x.\Delta p = h/2$. If one considers the uncertainties of the mean of several observations it may be somewhat less but of the same order, which we write $\Delta x.\Delta p \sim h$.

Thus, if one tries to make a very accurate measure of position, for example by using 'light' of very short wavelength, X-rays, or γ-rays, one automatically increases the uncertainty in the momentum of the system after the observation. What is gained in one is lost in the other.

This relation is truly symmetrical between position and momentum. To find the momentum one must provide some space in which the body can move, and within this space its position is uncertain. The more space you give it, the more accurately you can measure the momentum but the less accurately is its position known. As before, $\Delta x.\Delta p \sim h$. This is one example of complementarity. One can know *either* position *or* momentum as accurately as one likes by suitable experiments but *not* both. They are like the two faces of a coin; you may look at one or the other, not both at once. Note that momentum, not velocity, is the quantity complementary to position. If the object is very light, for example an electron, the uncertainty in velocity is correspondingly large. For a speck containing even a few atoms the velocity is very small and the principle is no real restriction. For example, if one had a cube of 1000 atoms of copper, and measured its position with an error equal to the side of the cube the uncertainty of velocity need only be about 300 cm/s, about 1 per cent the velocity of sound.

Since space and time are so closely connected (see 'Relativity', Chapter 2), it is not surprising that a similar result holds for time. The quantity associated with time T is energy E, and it can be shown that $\Delta E.\Delta T \sim h$.

This result leads to the remarkable conclusion that the most cherished principle of physics, the conservation of energy, is not always valid, or at least that there are occasions when it cannot be verified. This explains a strange difficulty that troubled Rutherford in the early 1920s.

By the same calculation that gives the law of scattering of α-rays and led to the discovery of the nucleus one can find how close an α-ray from uranium could get to the nucleus of another uranium atom if it chanced to hit exactly end-on (Fig. 3.5). It would then be at rest for an instant, retrace its path, and come out with its original velocity. One might expect the point of rest to be the edge of the nucleus from which uranium α-particles come. But in fact the α-particles used in the scattering experiments, which came from RaC and were much more energetic than those from uranium, penetrated considerably inside this distance and their scattering was still that calculated from a charged spherical nucleus. The true nucleus was therefore of radius smaller than the distance first found; but if so, why had the uranium α-particles acquired no extra energy in the strong repulsive field between the true nucleus and the distance first found?

The answer came independently from Gamow and from Condon and Gurney. The α-particle passes through a region, close to the nucleus, that is forbidden by the conservation of energy. This is possible on the indeterminacy principle because it would spend such a short time in it. Another way of dealing with the same problem is to consider the α-particles inside the nucleus as stationary waves of length h/p, where the momentum p can be calculated from the energy of the α-particles inside the nucleus. The region just outside the nucleus is a region of smaller energy. Work would have to be done to get *and hold* the α-particles there, but still farther away the energy will increase and becomes greater than that inside. By a calculation of a type similar to that successfully used to deal with the so-called 'total internal' reflection of light, Gamow and Condon showed that some waves will trickle over this 'potential barrier' and then (Fig. 3.14) decrease in wavelength as they reach regions where the energy of particles would be higher. These waves carry, on Born's principle, a finite probability of the presence of particles and so are the waves associated with the α-particle when emitted. The emission of these particles thus involves passage quickly through a region where, from the

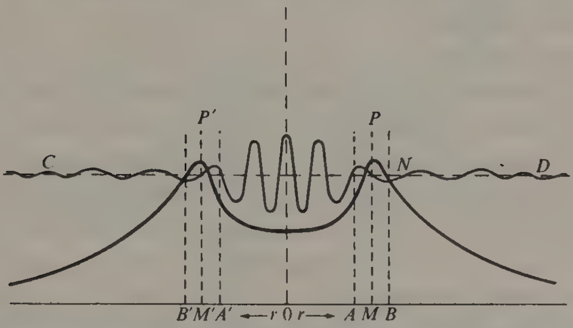


FIG. 3.14. Passage of waves through a 'forbidden' region. Heavy line shows potential energy of an α-particle in a nuclear field. Thin line represents wave-form, taking *CD* as base line and *AB*, *A′B′* as 'forbidden' regions. Bombarding αs will be scattered nearly normally unless they get closer to centre than *P*, which requires an initial energy greater than *PM* (i.e. *OQ* in Fig. 3·5 less than *OM*). *BN* is the energy gained by an α- particle *after* escape. *r* is the distance from the centre of the nucleus.

particle point of view, the law of the conservation of energy is violated. On the wave view this is merely a region with complex refractive index. Either point of view is perfectly tenable; the two never contradict.

For the interference fringes produced in electron diffraction, one cannot tell through what part of the crystal the particle passed, if one is interested in its path, except by measuring the sideways momentum accurately (Fig. 3.9). If this is done the uncertainty of the sideways *position* of the crystal is such that the fringes, whose position shifts with that of the crystal, would be blurred and invisible. If you find the particle you miss the wave, and vice versa.

Though the experiments are difficult it has been shown that both protons and atoms of helium show the same kind of wave properties as do electrons. There is good reason to suppose that the motion of the centre of gravity of *any* system would show these properties if it were possible to observe the extremely small wavelengths involved.

The particle properties of electrons were discovered before their wave properties. It was the reverse for light and for X-rays. The wave properties of X-rays were first shown by von Laue in 1913 by diffracting the rays by the atoms of a crystal; just as for electrons the three-dimensional array of these atoms acts as a diffraction grating, and the X-rays appear where waves of wavelength appropriate to the kind of X-ray used would be strongly diffracted. W. H. Bragg in 1914 used these waves to measure the wavelengths of the X-rays characteristic of certain atoms. His son, W. L. Bragg, almost simultaneously used the diffraction patterns of the crystals of rock salt (sodium chloride, NaCl) and sylvine (potassium chloride, KCl) to find the arrangement of atoms in these crystals.

A. H. Compton showed the particle properties of X-rays in a specially clear form in 1922. He found that some of the X-rays scattered from matter are more easily absorbed than the original ones, which means that their frequency is less and their wavelength longer (Fig. 3.15). This diminution of frequency he proved to correspond exactly in each case to the energy that a photon of energy $h\nu$ and momentum $h\nu/c$ would transfer to a free electron if it were deflected by collision with the electron into the direction of the scattered X-ray (Fig. 3.16). This energy comes from the photon. Here energy is conserved because the change of motion of the electron and of the frequency of the photon last indefinitely; so ΔT is infinite and ΔE zero for the process. The X-rays that have undergone Compton scattering are *not* capable of producing the Laue–Bragg diffraction effects. Their waves are said to be 'incoherent', that is, the phases of the waves of photons scattered from different electrons have no definite relation to

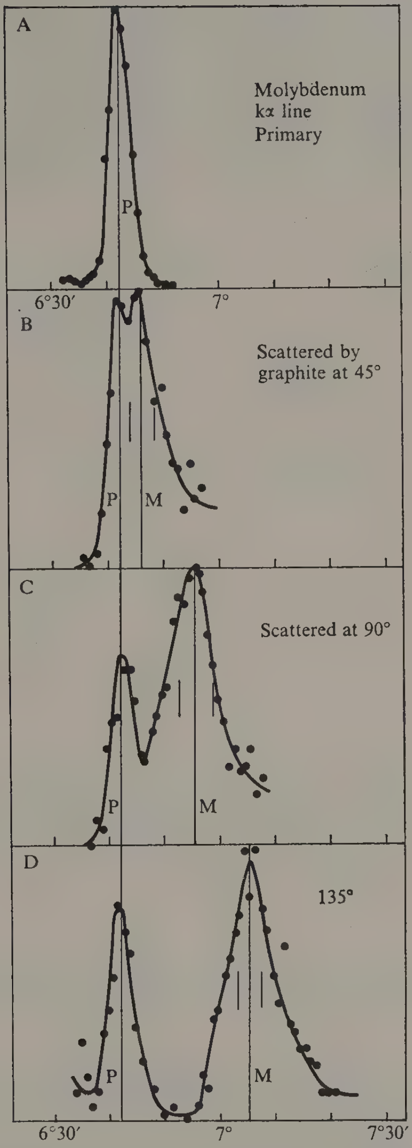


FIG. 3.15. Compton effect. Change of wave observed in scattered X-rays. Spectra of X-rays scattered by graphite at different angles, showing modified line wider than primary and displaced to theoretical position. From A. H. Compton and S. K. Allison, *The scattering of X-rays* (Van Nostrand, 1935).

one another and so cannot combine to give a diffracted beam. Observation of energy or momentum precludes observation of wave properties. The complementarity of waves and particles is a general truth.

It can be summed up in the following statements.

The concepts required for describing microphenomena differ fundamentally from those used for ordinary macrophenomena. The predictions of all valid theories are of probabilities, not of certainties. The character of the prediction depends more than before on the method used to observe the experiment, and this becomes an integral part of the experiment.

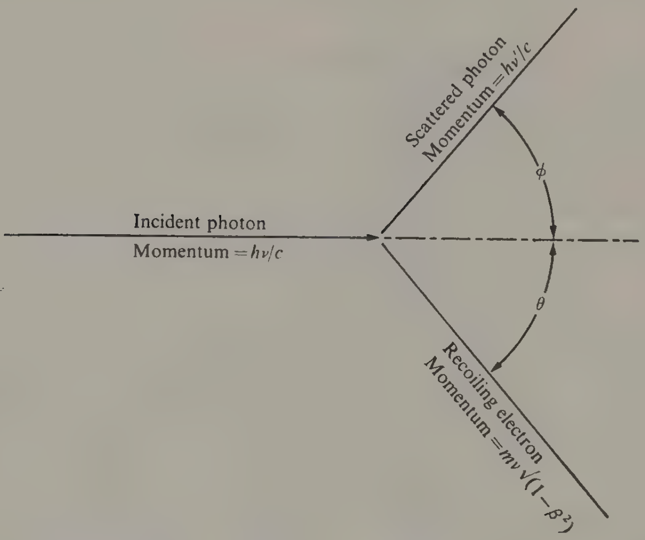


FIG. 3.16. Compton effect explained as a particle collision. When an X-ray photon is scattered by an electron at an angle ϕ the electron recoils at an angle θ, using some of the photon's energy and hence reducing its frequency. From A. H. Compton and S. K. Allison, *The scattering of X-rays.*

The attempt to observe one set of observables may interfere with the observation of another and in extreme cases quite preclude it.

The implications of these profound changes in physics go much beyond the mathematical processes by which the new theories can be translated into predictions. They involve a limitation on the rationality of nature. For long there have been principles asserting the impossibility of certain processes, for example the creation of energy; now we assert as a principle the impossibility of achieving certain kinds of knowledge.

Bound electrons

Electrons bound in an atom are regarded by the wave mechanics as held in a 'box' by the field of the nuclear charge. In the case of hydrogen to which Schrödinger's mathematics originally applied, there are certain special energies for which alone the wave is a steady vibration, with definite frequency proportional to the energy. These states correspond to Bohr's orbits, and the idea of a wave reducing to a steady vibration is the wallpaper simile (p. 69) turned into mathematics. As before, the energies of the steady states are the 'energy levels' of the atom; their difference divided by *h* is the frequency of the radiation emitted or absorbed as the atom changes from one state to another. But now this assumption is less arbitrary. If two states coexist, their different frequencies will make beats with each other as in acoustics and, also as in acoustics, the frequency of the beats will be the difference between the frequencies of the tones. The frequency of emission is the difference of the energy levels divided by *h*, as Bohr originally assumed. There is another acoustic parallel. Each steady vibration, except that of lowest energy, has certain surfaces at rest. These correspond to the nodes of a violin string, the nodal lines of a vibrating drum, or the nodal surfaces of a vibrating wind instrument.

In acoustics these nodal surfaces are determined by sets of numbers. In the simplest case, the violin string, there is only one number for each vibration, which is the integer greater by unity than the number of nodes. In Schrödinger's theory the corresponding three numbers are also integers. They are the quantum numbers. (Schrödinger's theory does not cover magnetic or relativistic effects, so there are only three quantum numbers (see p. 67).)

This gives a pictorial representation of a hydrogen atom. For other atoms with more than one electron the mathematics is much more complex and representation in space is no longer possible. In most cases the exact calculations become excessively difficult and approximations have to be used, which for many purposes are adequate.

The one other simple case is that of a gas of many electrons held in a box, which is a drastic but still useful simplication of the state of a metal. The electrons are the 'free' electrons whose motion under an electric field forms the current in the metal. The atoms of the metal that have supplied the electrons are of course positively charged and their averaged attraction prevents the escape of the electrons and can be regarded as turning the piece of metal into a box. Inside this box the electrons are supposed to exist as stationary waves (Fig. 3.17). Like waves of sound in a pipe, their wavelengths are fixed by the conditions at the ends, except that in the piece of metal transverse and oblique waves are allowed. In the pipe each wave is characterized by the number of its nodes, increasing from the fundamental to the successive overtones by one each time. In the metal, with its three dimensions, there are three sets of nodes, and any stationary vibration is determined by a set of three corresponding integers. These (Fig. 3.17) three numbers are the *quantum numbers* of the electron, and to them Pauli's principle applies; there can only be two electrons with any given set. Since the shorter its wave the higher is the energy of an electron, the long waves with low quantum numbers will be taken up first. The more electrons there are in the 'box' the higher the quantum number needed, and so the shorter the wavelengths that will have to be used. With one electron per atom to be accommodated, waves have to be used down to 2 or 3 × 10^{-8} cm with an energy of the order of 20 eV. This holds even at the absolute zero –273° C. At higher temperature a few electrons have wavelength shorter than needed by Pauli's principle and so more energy, but the contribution these make to the specific heat of the metal is very small. This work, of Sommerfeld removed a difficulty that had spoilt many earlier theories of metallic conduction, namely the apparent absence of any contribution of the free electrons to the specific heat.

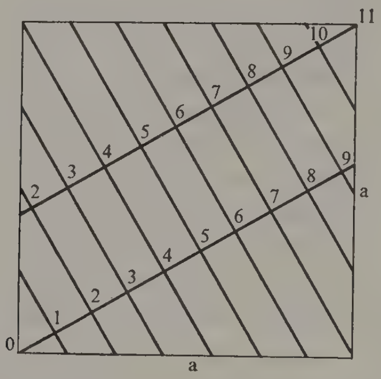


FIG. 3.17. Example of standing waves in two dimensions. The oblique lines are the crests and troughs of a possible set of standing waves in the square. They divide the vertical sides into four equal segments and the horizontal sides into seven. From M. Born, *Atomic physics* (Blackie, 1944).

The Dirac electron

The Schrödinger electron is non-relativistic and the wave function ψ is a scalar; so it could not replace the spinning electron where this was needed. Dirac (1928) put forward a theory that dealt with these difficulties and more. His electron has four wave functions, is relativistically invariant, and has an intrinsic magnetic moment and spin (moment of momentum).

Relativity led Dirac to the unexpected conclusion that states of negative energy are possible. After trying in vain to identify these with the proton, he came to the conclusion that they must normally all be filled and so unobservable. But if an electron were pulled out of such a state and given positive energy mc^2, the 'hole' left would appear as a particle of positive charge and mass equal to that of an ordinary electron.

These positive electrons, called 'positrons', were discovered in nature in the cosmic rays by Anderson and are emitted by some artificial radioactive substances (see p. 93).

Positrons disappear by recombination with negative electrons, which as it were fall into the holes. By relativity the energy thus set free is $2mc^2$, where m is the mass of either electron. It appears as two equal quanta of radiation in free space but near a massive particle, which can take up the recoil, a single quantum of the whole energy (about 1 million eV) can be emitted instead.

The reverse process is also possible, as Blackett and Chadwick have shown, *pairs* of electrons of opposite sign being produced by X-rays of more than 1 million eV.

This is the nearest man has ever got to the creation of matter, for the negative electron of the pair is indistinguishable from any other ordinary electron.

The existence of a particle bearing the same relation to a proton that a positron does to an electron has been proved experimentally by the use of particles accelerated to very high energies (about 2000 million eV). It has been suggested that some distant galaxies may be made of 'antimatter', positrons, and antiprotons.

Crystal structure

Since the discovery by V. Laue of X-ray diffraction and its use by Sir Lawrence Bragg to solve the structure of the simple crystals NaCl and KCl the science of crystallography has been completely transformed and with it our knowledge of the solid state. Something like 10 000 crystals have had their structure, that is to say the way in which atoms are packed inside them, determined, for perfect crystals are repetitive structures and if the arrangement in a unit cell is known all the rest follows.

As time has passed more and more complicated structures have been solved, the lead often being taken by W. L. Bragg and his pupils. In recent years very complicated organic compounds such as haemoglobin have yielded to modern techniques of analysis. The importance for biology, especially genetics, of this kind of attack can hardly be exaggerated.

The study of the solid state has not been confined to pure substances or perfect crystals. Very small degrees of impurity, which are often reckoned in parts per billion, may produce important changes in electrical properties, especially in semiconductors such as silicon and germanium. Most semiconductors in fact depend for their conducitivity on impurities. The wide application in technology of semiconductors in the form of transistors and otherwise has made the solid state the most important branch of physics as judged by the number of papers published.

Another branch of solid-state physics of great technical importance is the study of the mechanical imperfections of crystals, especially those known as dislocations. These are imperfections of various kinds concentrated on lines of strain running through the crystal. When the crystal is further strained the dislocations in it will move to relieve the stress but may be prevented from doing so by locking against one another. Most of the strength of metals is due to this action.

In brittle materials such as glass surface cracks are the more serious defects; they greatly magnify the applied stress. In metals this effect is retrieved by local plastic yielding.

It is, however, believed that crystals both of metals and of non-metals if entirely free from dislocations would be extremely strong; 10 times or more stronger in tension or shear than normal material. Very thin fibres of abnormal strength called 'whiskers' have already been made of certain substances, e.g. carbon, and there are hopes that if they can be improved, and if satisfactory methods of bonding them together can be developed, they may become of great technological value.

Isotopes

Chemistry is a property of the electrons in an atom, and mostly of the outer ones: if two atoms have different nuclei but the same nuclear charge, i.e. atomic number, they will have the same number of electrons and the same chemical properties, though the difference in the nuclei will give them different masses; this fact gives a possibility of separation by physical methods. For example, in a mixed gas with light and heavy molecules the kinetic energy being the same by equipartition the *velocity* of the lighter molecules will be greater and they will escape more rapidly through fine holes or pores. The difference, however, is never large and is very small for heavy atoms; so only a small degree of separation is produced by a single passage through a porous wall. To get good separation many passages are needed and the process is very tedious.

The existence of *isotopes*, as atoms with the same nuclear charge but different weight are called, was in fact established by Soddy. A number of cases were found in which two or more radioactive substances were chemically inseparable though their radioactive properties were quite different. This is easily understood in terms of the nuclear atom. The emission of an α-particle with a double positive charge will diminish the nuclear charge by two; the emission of an electron as a β-ray will *raise* it by one. Hence an α-particle followed by two electrons brings one back to the starting-point as far as charge is concerned, though the nucleus will be (about) four units lighter.

The first indication that ordinary non-radioactive elements might have atoms of different masses came from the study by J. J. Thomson of ‘positive rays’. These are formed under the same discharge conditions as cathode rays if a hole is made in the cathode, or better a tube passed through it, when they appear as light in the gas behind. By greatly reducing the gas pressure in this region, though keeping enough gas in front to maintain the discharge, Thomson was able to show that these rays consist of a variety of atoms and molecules derived from the gas in the discharge, mostly with a positive charge of one or more units; more rarely, and only in certain cases, with a single negative charge. This was done by separating the rays by superposed electric and magnetic fields, which sort them out according to their charge to mass ratio. They were recorded on a photographic plate as separate parabolic arcs (Pl. 3(b)). Among other gases tried was neon, an inert constituent of the atmosphere of atomic weight, as found from its density, of 20·2. This showed two parabolas corresponding to weights of roughly 20 and 22, always in the same proportion however the neon was prepared or purified. There were strong indications that both parabolas were due to atoms, which would accordingly be isotopes. Attempts made by Aston, who was working with Thomson, to separate them by diffusion through pipe-clay, showed some effect, but at the outbreak of the 1914 war the matter was still open to doubt. After the war Aston made some more attempts at separation, which were unsuccessful, but settled the matter by greatly improving the accuracy and resolving power of the electromagnetic separation. The consequences were surprising. First he proved that the two lines were indeed 20·0 and 22·0, giving the observed density of 20·2 (there was also a third faint one at 21). Then he found that a number of the common elements were mixtures of isotopes, notably chlorine, whose chemical atomic weight of 35·5 is due to a mixture of atoms of weights 35·0 and 37·0. In fact, single elements are the exception, but in some cases one isotope is much the commonest and so the others escape attention, as did those of hydrogen, carbon, and oxygen for some years after Aston's first discoveries.

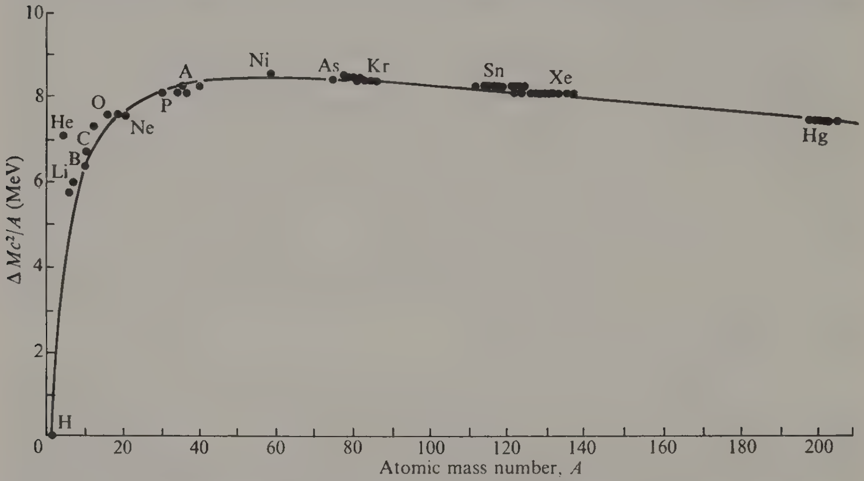


FIG. 3.18. Binding energy of some nuclei. *M* is the difference between the mass of an atom and the sum of the masses of the protons and neutrons it is supposed to contain. The ordinate is the energy per particle in MeV required to split the nucleus into protons and neutrons. The dots are experimental points from the measurements of Aston. From B. B. Rossi, *Cosmic rays* (McGraw-Hill, 1966).

Aston later improved his mass-spectrograph so that it would measure the small deviations from whole numbers† of the atomic weights of individual isotopes. These small deviations in mass are of great importance in seeing what conceivable reactions between nuclei would release energy, for by Einstein's law if the masses of the resulting nuclei do not add up to the masses of the initial ones the excess multiplied by the square of the velocity of light will appear as free energy (Fig. 3.18). This was recognized before nuclear energy on a large scale was thought possible.

† On the scale by which the commonest isotope of oxygen weighs 16·000.

Nuclear physics

In 1919 Rutherford proved that the collision of α-particles with nuclei of nitrogen could in rather rare cases cause the emission of protons with considerable energy (Fig. 3.19). This epoch-making discovery came from a study of the scattering of α-rays by hydrogen. Close collisions result in the proton from the hydrogen being given up to 8/5 times the original velocity of the α-particle, and a considerably longer range in air. By putting the detector (a scintillating screen) outside the range of the α-particles, it is thus possible to detect protons even though they are very few in number compared with the α-particles. Marsden (1914) found the expected protons, which Rutherford studied in detail after the war. The surprise came when he worked with substances supposed *not* to contain hydrogen, in particular with nitrogen. Long-range particles were then produced that were similar to those formed when hydrogen was used. However, the most careful exclusion of hydrogen failed to remove them. Though Rutherford, characteristically, grumbled at the apparent inability of chemists to make substances free of hydrogen, he soon made up his own mind that the new long-range particles did in fact come from nitrogen

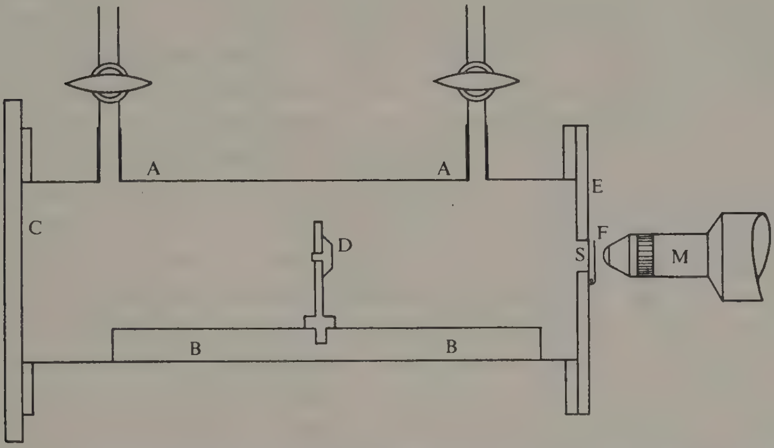


FIG. 3.19. Rutherford's apparatus for disintegrating nitrogen. The α-particles from the source *D* were absorbed in the gas. The opening *S* was covered by a thin sheet of silver. Outside the opening was a screen *F*. This showed scintillations when nitrogen, but not when oxygen or carbon, was in the box. From E. Rutherford, J. Chadwick, and C. D. Ellis, *Radiations from radioactive substances* (Cambridge University Press, 1930).

and were protons or just possibly atoms (then unknown) of mass 2. The early work in Manchester seemed to show that the range of the new particles was the same as that of those from hydrogen, but later work after his arrival in Cambridge showed that the former were more penetrating, thus removing all question of contamination. Magnetic deflection of the particles showed their mass to be 1 not 2. This brought the transmutation of the elements a stage further. Hitherto the only transmutations observed had been the spontaneous ones of radioactivity, a built-in process that many attempts had failed to increase or reduce by artificial means.

At first sight the new experiments suggested a splitting of the nucleus of nitrogen by the α-particle, but Blackett in 1925 showed that in fact the α-particle disappeared into the nucleus, which thus must increase in weight from 14 to 17. He did this by an improvement on the Wilson cloud-chamber method using simultaneous photographs from two directions so that a reproduction in three dimensions became possible. He took about 23 000 pairs of photographs containing over 400 000 tracks, among which were 8 events showing the production of a proton (Pl. 4).

Rutherford's later work showed that all elements up to and including phosphorus give protons, except helium, carbon, and oxygen. It was *prima facie* unlikely that the heavy elements would do so as their large nuclear charges would not allow the α-rays to approach sufficiently near. However, α-rays are not the only possible bombarding particles, and the history of physics up to World War II is one of the development of nuclear bombardment by artificially accelerated atoms, leading to a vast development of nuclear chemistry and spectroscopy.

The pioneers in this field were Cockcroft and Walton. They were influenced by the theory of Gamow (see above). On this theory a charged particle approaching a nucleus directly would have some chance of leaking through the potential barriers caused by the nuclear charge even though its energy was insufficient to go over the barrier on a conventional calculation.

Cockcroft and Walton's early apparatus (Pl. 5) could accelerate protons up to 500 kV. The protons were obtained from a discharge tube and then accelerated in a high vacuum. In the first experiments Cockcroft and Walton looked for γ-rays from targets of beryllium and of lead and did not find them at 280 kV. In later experiments (1932) they found α-particles at even lower voltages, the yield increasing rapidly with voltage. They supposed, correctly, that the proton entered the nucleus of ^{7}Li and that the compound nucleus split into two α-particles with conservation of charge and a loss of mass, which agreed well with the observed energy of the α-particles and the exact masses of the atoms concerned as measured by mass spectroscopy. Boron also gave a copious effect, the nucleus of ^{11}B splitting into three α-particles. The α-particles were detected by scintillations, but the methods of disintegration were confirmed by observing the tracks of the particles in a cloud chamber (Dee and Walton 1934). These effects were in a sense the reverse of Rutherford's, protons giving α-particles instead of α-particles giving protons. Very soon after Cockcroft and Walton's work the recently discovered 'heavy hydrogen' of mass 2, now called deuterium, was used in the U.S.A. Its nucleus, the deuteron, has considerable energy. Collisions between two deuterons give nuclear reactions yielding energetic particles at very low impact energies because of the relatively small repulsion of the singly charge nuclei (Oliphant, Horlick, and Rutherford 1934). These reactions are used in the hydrogen bomb and are the basis of the hope that nuclear energy released by great heat can be used for peaceful purposes.

These early experiments were the precursors of a vast amount of work on nuclear reactions, which is still going on. It has been greatly helped by the discovery of novel methods of accelerating charged particles. Some of these, such as the cyclotron, invented by Lawrence, and the betatron for electrons by Kerst, depend on the particle to be accelerated going round many times in a circle. The van der Graaf generator uses straight accelerating fields produced by the mechanical transport of electrostatic charges. The linear accelerator has a straight line of cavities each with its alternating axial field with which the particle must keep step.

An even more important way of producing nuclear change came as a by-product of Chadwick's great discovery of the neutron in the '*annus mirabilis*' of 1932. It had been known for some time that boron and beryllium when bombarded by α-rays gave in addition to protons a very penetrating radiation. Further, this radiation produced secondary radiation in substances containing hydrogen and to a lesser extent nitrogen. Attempts had been made by the Joliot-Curies and others to estimate the quantum of this radiation supposing it to be a γ-ray, and the secondary radiation in hydrogen to be protons recoiling from a scattering collision, but this led to contradictions. Chadwick came to the conclusion that the radiation was probably material and from comparison of the effects of secondary particles from hydrogen and nitrogen was able to deduce the mass of the primary particle that came out to be about that of the proton. Rutherford many years before had predicted the existence of such a neutron, but attempts to find it had so far always failed.

A number of its properties were readily established. Having no charge, its interaction with the electrons of atoms is negligible. Only when it struck a nucleus could it be scattered or absorbed, but because it is not repelled by the nuclear charge nuclear reactions are easy. It is scattered most by substances rich in hydrogen (i.e. in protons), for a collision with a proton of mass roughly equal to its own deflects and slows it down. There are also nuclei, such as boron, with which it reacts very readily; in such cases the reaction between the neutron and the nucleus is favoured by a process that has much in common with mechanical 'resonance' between vibrating systems and is called by the same name.

There are many kinds of reactions between neutrons and nuclei. Some result merely in scattering of the neutron, with or without loss of energy. Others involve the capture of the neutron to form an isotope of the original nucleus. This new nucleus may be stable or it may break up. If it breaks up, the break may occur almost at once or after a delay that may last for years. In all kinds of reaction the process may be accompanied by γ-rays. The break may be of many kinds, ranging from a complete break into two parts of comparable size, known as fission, the basis of nuclear reactors, to the emission of a particle of electronic mass, usually a positron, or of a γ-ray only. Many of these reactions were studied by Fermi in the middle thirties. He used 'slow' neutrons that had been reduced to thermal equilibrium by repeated collisions with protons in water or paraffin surrounding the source of neutrons. Though some reactions possible with neutrons of, say, a million electron volts cannot happen with slow neutrons the majority happen more readily with them. One may visualize the slow neutron as staying longer in the vicinity of the nucleus and so more likely to react with it. The variety of processes by which a nucleus can get rid of excess energy after a collision is enormous and increases with increasing energy and complexity of the atomic projectile that causes it. Ultimately, perhaps after several transformations, the system will end up as one or more of the known stable nuclei.

In the early days of nuclear theory it was supposed that nuclei were made of protons and electrons only. This led to difficulties; for example, to find room for the electrons in the nuclei whose radii were known approximately by α-ray scattering and are a small multiple of 10^{-13} cm. The discovery of the neutron cleared this up and nuclei are now supposed to be made of protons and neutrons. For a nucleus of atomic number Z and atomic weight A, Z protons and A–Z neutrons are needed. Isotopes have the same Z and different A. As Fig. 3.20 shows, the stable isotopes lie in a fairly narrow strip if A is plotted against Z. Nuclei of too few neutrons for their Z tend to emit positrons to reduce Z and so get back into the region of stability; those with too many may emit ordinary electrons, turning one or more neutrons into protons. Indeed, the free neutron itself does this, splitting up with a half-life of 12 minutes. Some unstable nuclei, instead of emitting a positron, capture an electron from an inner level of the atom.

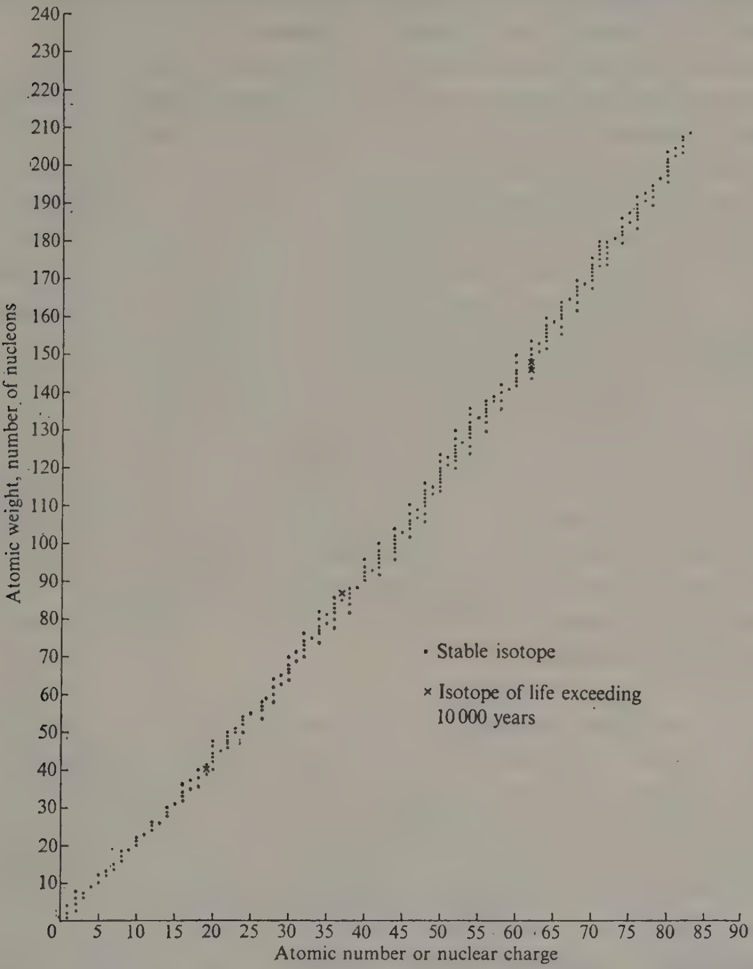


FIG. 3.20. Stable and long-life isotopes. · denotes stable isotope. × denotes isotope of life exceeding 10 000 years. Missing elements 43, 61, 84, 85, 86, 87, 88, 89 have only radioactive isotopes of shorter life. The long-life isotopes of thorium (90), proactinium (90), and uranium (92) have been omitted.

Artificial radioactivity of the β-type, i.e. positrons or electrons, was discovered by the Joliot-Curies in 1934. They found that several of the lighter elements when bombarded by α-rays emit β-rays of one sign or the other with exponential decay like that of natural β-ray emitters. It is rather surprising that this discovery was so long delayed.

The mechanism of β-ray emission, both neutral and artificial, is more obscure than that of α-rays. It involves what are technically called 'weak interactions' (as opposed to the 'strong interactions' that hold nuclei together against the natural electrical repulsion of their protons) and the ordinary electromagnetic forces that govern the electrons outside the nucleus. If the β-rays from a narrow source are bent by a strong magnetic field they form the analogue of an optical spectrum. This appears as a continuous spectrum with a sharp upper limit of energy. There may also be lines sharply defined in energy; these are of secondary origin. The other electrons present a problem. It was proved that one, and only one, was produced for each atom disintegrating. Why then are they of different energies? The simplest explanation would be that they start with the maximum energy and mostly lose a variable amount before being recorded. Ellis and Wooster disproved this in 1927 by measuring in a calorimeter, which absorbed all the γ-rays, the total energy produced. This corresponded well with the *mean* energy, of the β-rays observed. But if different atoms really did emit different amounts of β-ray energy, why did the resulting atoms, themselves often radioactive, behave so normally in future decays; for they must have very different amounts of energy still left in them?

Pauli in 1934 put forward the theory that a new very penetrating particle must be emitted at the same time as the electron with varying energy so that the sum of the two energies is constant. Fermi adopted the idea and named the particle the neutrino. Like a photon it has zero rest mass and no charge; it has enormous penetrating power, and would easily go through the whole earth. At first the suggestion was received with hesitation. It is too easy to imagine entities whose properties defy observation and the fate of the undetectable ether was recent enough to act as an awful example. However, the neutrino has lasted and small effects have accumulated. For example, in electron capture by the nucleus ^{7}Be which transforms in this way to ^{7}Li, Allen in 1942 found a small recoil corresponding to the momentum of the neutrino required. Here there is no electronic momentum to upset things. More recently, individual neutrinos† have been detected near a nuclear reactor by their action in allowing protons to turn into a positron and a neutron. The neutrino, and its antiparticle, the antineutrino, the first emitted in positron

† Strictly speaking these were antineutrinos.

β-decays, the second in electron β-decays are now quite generally accepted.

Our knowledge of the nucleus is still very incomplete in spite of masses of evidence on nuclear levels. Transitions between these give γ-rays, as transition between electron levels in atoms give photons. Clearly there must be a nuclear force to keep the nucleus from bursting under the mutual repulsion of the protons, but its nature is still unknown. It is believed to be the same between any pair of particles, proton or neutron, together called *nucleons*. It is a short-range force and, unlike gravitation and electrostatic forces, but like chemical forces, can be saturated. A large nucleus has, in general, no more attraction for an extra neutron than a small one. Chemical forces are understood in principle and depend on the possibility of the complete or partial exchange of electrons between atoms. Something similar is probable for nuclear forces. In the study of nuclei it has been generally assumed that concepts of the *kinds* that have proved valid in the arrangement of electrons in atoms also apply, and no proof to the contrary has been found. There is evidence for something analogous to the arrangement of electrons in levels determined by quantum numbers. Some numbers of particle seem specially stable, an analogy to the inert gases. But the knowledge obtained about principles, as contrasted with individual facts, in the last thirty years is disappointing when one considers the great expenditure of effort and money on the subject.

Nuclear Fission

The most startling discovery in this period, and the best known, was that of nuclear fission by Frisch and Meitner at the turn of the year 1938–9 following work by Hahn. Both fast and slow neutrons can be captured by the nucleus of uranium, leading to a division of it into two nearly equal parts. This process releases about 200 million eV of energy (200 MeV) in contrast to ordinary radioactive processes of a few MeV; and, most important, a few neutrons are emitted at the same time, thus suggesting the possibility of a chain-reaction the rate of which would increase exponentially unless checked, with immense emission of energy.

The existence of vast stores of energy in the atoms of uranium and thorium is implicit in Rutherford and Soddy's theory of radioactivity (p. 53). Aston's measurement of the masses of individual isotopes showed that the heavy nuclei contain progressively more energy per atom as the atomic weight increases above about that of iron. Thus nuclei heavy enough to split into two pieces, each substantially heavier than iron, are *prima facie* able to yield energy by doing so. It may not, of course, be possible to make them split as desired. Even in some of Rutherford's original experiments with light atoms, an α-particle freed more than its own energy by releasing a proton. But since only a few per million of the α-particles released a proton and the rest wasted their energy in heat there was no hope of a useful source of energy here. The neutron increased the chances, for neutrons react only with nuclei, and may do so even when they have lost kinetic energy, but to make them multiply was difficult. The reactions that gave two neutrons for one worked only with fast neutrons. Fission altered the situation. The number of neutrons produced per fission was about three, and slow neutrons could produce fission. However, there are ways in which neutrons can be lost even in reactions with uranium, and it was not quite certain that a chain-reaction could be achieved till Fermi actually did so with a 'pile' of uranium and pure graphite at Stagg Field of the University of Chicago on 2 December 1942. The extra difficulties involved in making a bomb, which means making a chain-reaction work with fast neutrons, are matters of technology outside the scope of this book.

The importance of fission is greater in technology than in science. It is of interest as a case in which an approximate theory of the nucleus yields very useful results. Bohr and Wheeler (1939) showed that the uranium nucleus behaves in a way analagous to a liquid drop held together by surface tension. The nuclei of uranium contain over 230 protons and neutrons, a number large enough to make a continuous medium a reasonable approximation. The actual splitting force is supplied by the mutual repulsion of the protons. When a neutron comes in the nucleus is set into violent vibration, which may make it unstable.

Cosmic rays

The careful investigation of a small outstanding discrepancy has sometimes led to important discoveries in physics, but none so great as those that have come from the discovery of cosmic rays. X-rays led at once to a great interest in all aspects of the conduction of electricity through gases, including gases in their natural state—not that this was a new subject. Coulomb in 1785 had come to the conclusion that the small leakage from a charged body suspended in the air could not be accounted for by leakage over the strings. Much later, the almost simultaneous experiments of Geitel and Wilson (1900) were the first to make definite progress. Wilson found that the rate of leak increased with the size of the vessel. This suggested a volume ionization. Most of this was due to traces of radioactivity in the apparatus itself, to penetrating radiation from the materials of the building, and to radioactive gases diffusing into the air from the soil. However, there seemed to be a residue. At the time it was considered possible that all or many ordinary atoms might be slightly radioactive.

Various experiments had been made at moderate heights which showed that the ionization diminished as one got further from the radioactivity in the ground. Hess in 1911 found from observations in a balloon that, though the ionization diminished with height at first, it then increased again to more than the value at the ground, indicating that there was a radiation from outside. Further work, notably by Kohlhörster and Millikan, fully confirmed this. The new radiation was fairly constant with time, was much stronger at great altitudes, and its absorption in air, assuming it came from outside, was about equal to that in an equal mass of water. Millikan believed that the radiation was, like the γ-rays, in the form of photons and derived its energy from the annihilation of matter in distant space. We now know that the matter is very complicated indeed, that it involves types of particles unknown to Millikan, and that few if any of the primary rays are photons. We still do not know where they come from or how they get their energy. It would take too long to explain how the present conclusions were reached but before setting them out I must briefly describe the various methods used and the information they can give. The early experiments measured the leak from an insulated charged conductor. Next came the Geiger–Muller counter. This is a gas-filled tube with a wire down the middle and a potential between the two, nearly but not quite enough to produce a discharge. When ions are produced by the passage of a ray the discharge is triggered, is automatically recorded, and the counter reset. By using two or more counters and recording events only when a certain selection go off at once one can detect rays going in a particular direction only: a cosmic-ray telescope. Then came the application of the Wilson chamber, with a magnetic field due to a large magnet, by which Anderson discovered the positron. The magnet bends the tracks of charged particles—only charged particles in fact ionise and so cause tracks. If the direction of motion is known, the *sign* of the charge follows from the sign of the curvature, i.e. to left or to right. The radius of curvature is in proportion to the momentum of the particle (assumed to be of unit charge) and varies inversely as the field. If the track is not too dense it may be possible to count the drops, i.e. the number of ions; this gives the speed, provided it is not too near that of light, again assuming unit charge. Multiply-charged particles usually have very dense tracks.

Blackett made the cloud chamber more effective for cosmic rays by combining it with counters (Fig. 3.21). Only when the counters fired was an expansion made and stereoscopic photographs taken, thus enormously reducing the waste of plates and time in scanning them. Finally Powell greatly developed the discovery of the Viennese ladies Blau and Wambacher that photographical plates record nuclear disintegration by cosmic rays like a cloud chamber, the plates being scanned under a microscope (usually after a suitable exposure at a great height). Because of the small scale, magnetic fields produce no appreciable effects; otherwise, plates can do nearly all that the chamber can with far less apparatus but at greater cost for scanning, at least until mechanical methods of doing this were devised. The sensitive film is 3000 times denser than the air in the chamber; more of the track can therefore be seen, often including its end (Pl. 6).

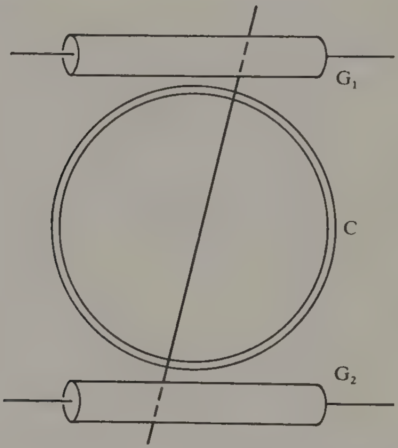


FIG. 3.21. Counter-controlled cloud chamber. A cosmic-ray particle passing through the two Geiger–Muller counters G_1 and G_2 and through the cloud chamber C produces a coincidence. The signal pulse from the coincidence circuit promptly triggers the expansion of the chamber before the ion pairs left by the particle have time to diffuse away. Vapour condensation around the ions produces a visible trail of droplets along the track of the particle. From B. B. Rossi, *Cosmic rays* (McGraw-Hill, 1966).

Primary cosmic rays are composed of the nuclei of the lighter atoms. Over 90 per cent are protons and most of the rest helium (i.e. α-rays), but some are heavier nuclei, up to iron and very rarely even beyond. Their energies run from about 10^9 eV, below which it is difficult for them to reach the atmosphere through the earth's magnetic field, to an extreme upper limit of 10^{20} eV, which exceeds by a *factor* of a hundred million the highest energy yet produced for an artificial accelerator. Energies anywhere near this upper limit are, however, excessively rare; one may literally have to wait for weeks before getting one. The *total* number of primary particles is of the order of one per second for each 2 cm^2 of surface.

When the primaries enter the atmosphere they collide with the molecules of air. At these high speeds the ionization produced is slight and the important effect is nuclear collisions. These result in a more or less complete break of the nucleus struck. The proton, if it is a proton, may keep a large part of its energy; some of the rest is shared between the protons and neutrons of the nucleus struck, which separate with considerable velocities. Part, however, may go in the production of a new type of particle: *mesons*. These particles are thought to be concerned in the mechanism of nuclear forces (see p. 91). Mesons, of which there are several kinds, are all very ephemeral bodies. Most have a single charge of either sign, but some are neutral. Those of negative charge may enter the nucleus of another atom. They then give up the very considerable energy (~100 MeV) derived from their mass, plus any kinetic energy they may have, which heats the nucleus of the new atom so that a number of neutrons and protons 'boil off' (Pl. 7(a)). If the atom is a light one the nucleus may be entirely dispersed. Alpha particles also are common among the fragments.

Of the mesons shown in Table 3.1 π-mesons, now usually called *pions*, are the commonest. Unless absorbed by a nucleus, they have a half-life of $2{\cdot}5 \times 10^{-8}$ s and then they turn into bodies originally called μ- (mu) mesons, now *muons*. Though it was these latter particles to which the name 'meson' was originally given, and which were the first to be found, it has now been decided that they are more akin to electrons in spite of the great difference in mass. However, they are the bad boys among these new particles and refuse to accommodate themselves to the theories.

When a charged pion comes to rest, slowed down by the ionization it caused, it transforms itself into a muon. No other track is seen in the photographic emulsion but the tracks are not in line; so to balance

TABLE 3.1

Mass of electron taken as unit, charge of electron as -1

Leptons	*Baryons*	*Non-conserved particles*
(spins all ½)	(spins ½ or $\frac{3}{2}$)	
Neutrino v *Antineutrino v* Mass zero, stable (perhaps two kinds of each)	*Proton p* charge +1 *Antiproton* $\bar{p}$ charge −1 Stable, except for mutual annihilation Mass 1836	*Quanta* Zero rest mass stable *π-Mesons* (pion) π^+, π^- (anti-bodies) Lifetime $2{\cdot}6 \times 10^{-8}$ s Mass 273
Electron e^- stable *Positron* e^+ charge +1 stable, except for mutual annihilation with electron	*Neutron n* Lifetime 12 min Mass 1838 Also anti-body $\bar{n}$ Lifetime 12 min	π° lifetime $\sim 10^{-16}$ s Mass 264. Is its own anti-body *K-Mesons* (kaon) K^- lifetime $1{\cdot}2 \times 10^{-8}$ s
mu-mesons μ^+, μ^- (anti-bodies) Lifetime 2×10^{-6} s Mass 207	Λ° neutral $\bar{\Lambda}^\circ$ antibody Lifetime $2{\cdot}7 \times 10^{-10}$ s Mass 2182	Mass 966 K^+ anti-body to K^- K° anti-body to K° Lifetime 10^{-1} and 10^{-10} s (The decay of these bodies is peculiar) Mass 975
	Σ^- lifetime $1{\cdot}7 \times 10^{-10}$ s Mass 2340 Σ° lifetime $< 10^{-13}$ s Mass 2326 Σ^+ lifetime 9×10^{-11} s Mass 2327	
	Ξ^- lifetime $\sim 10^{-9}$ s Mass 2582 Ξ° lifetime 10^{-7} s	
	Ω^- lifetime $\sim 10^{-10}$ s Mass 3300	

momentum another particle is needed that does not ionize. Its rest-mass can be shown to be small and it is assumed to be a neutrino. When the whole of the muon track can be seen it is always of constant length. This is consistent with the idea of a constant release of energy of 4 MeV and a single invisible particle. But when the muon decays in turn, which it does in about 2×10^{-6} s, an electron comes out with *variable* energy. Conservation of momentum and energy now require two invisible particles, presumably neutrinos. It has been surmised that the neutrinos associated with muons are different from those produced in ordinary β-decay.

Muons react very weakly with nuclei. Negative muons with their relatively long life must spend the latter part of it circulating round the nucleus of some atom; but unless it is a fairly heavy one their decay is not hastened.

Early in the study of cosmic rays it was found that those at ground level are not absorbed exponentially. If blocks of lead are placed above the detector the first few centimetres of thickness produce a rapid decline to about 75 per cent of the original effect; after this further thicknesses of lead have diminishing effects. Using the language of X-rays, there is a 'soft' 25 per cent and a 'hard' 75 per cent of the radiation, though the division cannot be made precise with this kind of experiment.

At ground level, where the primary particles have all disappeared, the 'hard' component is mostly muons. The 'soft' component is a mixture of electrons and photons forming what are known as 'cascades'. The process is as follows. Fast muons decay into electrons. These in turn, by passage through atoms, produce photons, just as cathode rays produce X-rays when they hit the anticathode, but the energy is much higher, well above the million eV required to make an electron–positron pair. Other high-energy photons come from the decay of the unchanged pions (see Table 3.1). Most of the photon energy goes into the production of high-energy pairs, but sometimes photons make knock-on electrons by Compton effect (p. 76). All these electrons in turn collide with atoms and make a second generation of photons. The whole forms a 'cascade' (Pl. 8). Each generation has more particles but of less average energy, until this is too low to produce pairs when the process stops. In this way the comparatively few particles of high energy may indirectly cause a substantial fraction of the ionization at ground level.

It may be wondered how it is possible to measure such extremely small lifetimes as 10^{-10} s, but in terms of *distance* it is easy. At one-tenth the velocity of light, which is only 50 000 eV for a muon, 10^{-10} s is 3 mm of track, a long distance on a plate. The three unknowns of a track, mass, velocity, and charge, must be determined from ionization density, range (if the track stops in the emulsion or chamber), curvature (if in a chamber with magnetic field), and scattering as measured by small kinks in the track. Multiple charges seldom occur except when a nucleus has been broken up; the track, being much denser, can then usually, but not always, be distinguished (Pl. 7(b)). Theoretically measurement of any two of the four quantities listed could determine the velocity and mass, assuming unit charge, but the laws governing the variation of the quantities measured with the velocity and mass of the particle are such that the unavoidable errors often lead to great uncertainty in the answers.

The sources of cosmic rays are still a matter of dispute. It is little use looking for their direction with counters arranged as a telescope, for the earth's magnetic field is too upsetting. More, indeed, come from the west than from the east but all this proves is that the majority have a positive charge. The intensity near the (magnetic) equator is slightly less than at high latitude. This shows only that some at least are charged. Solar disturbances, such as flares and sunspots, affect the measured intensities, more when the measurements favour low-energy particles, but to some extent even for those of higher energy. However, this does *not* prove that many of the particles normally come from the Sun; so large a proportion of those of low and medium energy are shielded from the Earth by its magnetic field that a change in this may be more important than a change in the initial flux of particles. The small magnetic variations measured at the Earth's surface are certainly related to disturbances in the Sun, and may be larger farther from Earth. However, it seems reasonably certain that the very large extra number of the more absorbable rays sometimes observed in solar flares—notably that of 26 February 1956—are actually particles coming from the Sun. But not all cosmic rays, not even probably a large fraction, come from the Sun. For one thing, the most energetic particles would only be slightly bent on the way, yet show no preference for this direction. It is still not even certain whether all the rays come from our own galaxy or from remoter space.

Many astronomical bodies show magnetic effects that could accelerate charged particles. Fermi suggested a method by which their energies could be increased, namely by collisions with regions of space containing ionized gas laced with magnetic lines of force from which the particles would rebound with, on average, increased energy. If the cosmic rays come from stars, these must be much better emitters than is the Sun; otherwise, the Sun because of its nearness would dominate as much in cosmic rays as it does in light.

Fundamental particles

These misnamed entities are, with the exception of electrons, protons, and presumably neutrinos, all decidedly ephemeral (see Table 3.1). They all decay, sometimes in several stages, and finish up as the three above, often with a balance of energy in the form of photons. Their study has been for about twenty years the most exciting part of physics, not only because of the strangeness of the phenomena but because it seems likely that if their relationships to one another were fully understood the problem of nuclear forces would be solved. Probably more money has been spent on apparatus mainly intended for this purpose than on all other research on physics since the time of Archimedes. Progress has undoubtedly been made, but there is still quite a way to go.

Besides the particles described above, several more were initially discovered in cosmic rays: the *K*-particles (originally called tau mesons) by the photographic techniques (Bristol and Imperial College), and the Λ (lambda), Σ (sigma), and Ξ^- (xi minus) particles by Rochester and Butler at Manchester. In recent years, however, since it has become possible to produce these particles by the artificial acceleration of protons and of electrons giving photons of very high energy, the use of cosmic rays for their study has become less important. The convenience of a controllable and intense source outweighs the disadvantage of the cost and complexity of the machines. Much of the information in Table 3.1 has been obtained thus. The important Ω^- (omega minus) and Ξ_0, besides a large number of what are called 'resonances', have actually been discovered by machines. Resonances are extremely short-lived states that cannot actually be observed as particles with visible tracks.

In this bewildering mass of data, even so probably incomplete, a few conservation laws can be found. Electric charge is conserved. When a charged particle disappears, either another of the same charge takes its place or one of opposite charge disappears, and conversely if a new one appears. More recently it has been established that there is a class of light particles called 'leptons', which is formed of electrons, muons, and neutrinos (probably of two kinds). Each has its 'antibody', charged in the first two cases, neutral in the last. Body and antibody can be created together and can annihilate one another, as we have seen with electrons and positrons. When a muon turns into an electron it emits two neutrinos as stated above. To satisfy the conservation rule these two must be a neutrino and an antineutrino.† All leptons have a spin of $\frac{1}{2}$ in quantum units, and the spins combine according to quantum rules which for two spins only means adding or subtracting to give 1 or 0. Particles with spin $\frac{1}{2}$ are called *fermions* and obey the Pauli principle.

Another group whose number is conserved are named *baryons*. These

† Possibly there are two kinds of neutrinos, each with its antineutrino, one set concerned with the muon, the other with the electron.

are all at least as heavy as the proton. They are: proton, neutron, lambda, sigma, xi, and omega particles. The neutron and lambda are unchanged but are believed to have antiparticles. The antineutron would disintegrate to give an antiproton and positron; it and the antiproton have been observed. Baryons disappear, like positrons, by combining with the normal corresponding particle but with much greater evolution of energy because of their greater mass.

The class of mesons, namely pions and *K*-mesons, like photons, are *not* conserved in numbers, though their energy is. A photon, for example, can raise an atom to an excited state and disappear in the process. The atom will re-emit the energy but it may be in several quanta whose energies add up to that of the original photon. Photons have spin unity, the others have zero spin: called *bosons*, they do not follow the Pauli rule. One must distinguish between conservation and stability. The photon is *not* conserved, but is stable, as is proved by the fact that distant starlight reaches the earth after many million years. The neutron is *unstable* when free but conserved as a baryon, since it yields a proton which is one also.

Recently considerable progress has been made in the study of baryons by considerations of symmetry; these have led to the discovery of the Ω-particle. The types of symmetry considered are not spatial only but may involve the charge.

On the other hand, considerations of symmetry do not always work among the leptons. Thus, Yang and Lee proposed experiments that have disproved the long-accepted view that two sets of apparatus that differed only in being mirror-images of each other would necessarily produce the same result, which is not true if part of the 'apparatus' is a nucleus having a β-decay. A negative β-particle has a twist like a left-hand screw; a a positron a twist like a right-hand screw; the accompanying neutrinos go reversely.

Low temperatures

The history of the study of very low temperatures since 1900 is a remarkable example of completely unexpected discoveries coming from an apparently routine line of research. By 1900 all gases, with the exception of helium, had been liquified. In July 1908 Kamerlingh Onnes at Leiden succeeded in liquifying helium by precooling the gas in liquid hydrogen and then expanding it through a nozzle, which, if the initial temperature is low enough, cools it, and making the process regenerative by using the exit gas to cool that approaching the nozzle. By evaporating liquid helium Onnes was able to get within about 1° of absolute zero (0° Kelvin), but the helium did not solidify. It is now known that it requires 25 atmospheres pressure to make it do so.

In April 1911 Onnes discovered that mercury when cooled in liquid helium had an electrical resistance too small to measure. This was not unexpected, and it took two further years of work before it was clear that he had not confirmed his theory of electrical resistance but instead had made one of the most startling discoveries ever made in physics. He called it superconductivity. It is only after fifty years of intensive study that a fairly satisfactory explanation for this phenomenon has been achieved.

The principal facts are now known to be as follows. Most pure metals, but apparently not all, and certain intermetalic compounds, for example niobium–tin, at a temperature characteristic of the substance but below about 20°K (10°K for pure metals) completely lose all resistance, so much so that a current, once started, will continue apparently undiminished for years with no battery to maintain it. The effect depends upon the state of the material; thus ordinary tin is superconductive, but the hexagonal modification grey tin is not. The superconductive current is destroyed at once either by heating the metal above the characteristic temperature or by applying a magnetic field; the field needed increases from zero at the characteristic temperature to a value characteristic of the metal but of the order of 100 gauss at the lowest temperature obtainable. Since currents produce magnetic fields this automatically limits the current a given piece of metal can carry without destroying its superconductivity.

Metals and metallic compounds behave differently in relation to magnetic fields. Superconducting metals put in a field too weak to destroy the superconductivity acquire a surface current that cancels the field inside. This is a natural consequence of low resistance, the only difference from normal metals being that the surface current persists. If the metal is now heated above the critical temperature it becomes normal and the magnetic field enters as expected. If, however, this process is done reversely and a metal already in a magnetic field is cooled until it becomes superconducting the field is *not* retained but expelled. When the external field is removed, no currents remain on the surface, either while the metal is superconducting or when it is heated above the critical temperature. The metal not only refuses to let magnetic induction in but expels it if already there when it is made superconducting. This is called the Meissner effect. In one respect the behaviour of a superconducting metal is more symmetrical than in the normal case. A steady current in a normal metal produces a steady magnetic field, but only a *varying* magnetic field causes a current. In a superconductor current and magnetic field are simply proportional. The behaviour of metallic compounds is less clear-cut. Magnetic flux can penetrate these substances with relatively low fields, but it does not destroy the superconductivity; this only happens when a considerable field has penetrated. Obviously these 'class II' superconductors are most important for possible practical uses, since perfect conductors will be of technological value only if they can carry considerable currents, that is, exist in considerable magnetic fields.

The explanation seems to be bound up in the finite thickness of the surface currents on superconductors, believed to be about 10^{-5} cm, i.e. a few hundred unit cells of the crystals. There are thermodynamical and experimental reasons for thinking that the critical magnetic fields for wires as thin as this should be greater than for thicker ones and it is supposed that 'class II' semi-conductors contain a great number of thin conducting channels. Already superconducting electromagnets are being wound with wires of substances such as niobium–tin, which can generate fields up to 100 000 oersteds without, of course, any resistance.

Superconductivity depends on an interaction between the electrons and the lattice vibrations, since the transition temperature is different for specimens made of different isotopes of the same element whose different masses cause different lattice frequencies but little other difference. As a result of the work of Cooper, Bardeen, Schreiffer, and others it seems likely that pairs of electrons are involved that have the same velocity but in opposite directions and opposite spin. The persistent superconductivity current may be analogous to the persistent orbit of an electron round the nucleus.

There exists a second surprising phenomenon called superfluidity, clearly allied to the first but strikingly different in that it exists only in one substance, namely liquid helium; moreover, the helium must be the common isotope of mass 4, not the rare one of mass 3.

It was discovered at Leyden in 1930 that liquid helium shows an abnormality in thermal behaviour at 2·2°K (called the lambda point) resembling that at a change of state. It was years before helium's strange behaviour below this point was straightened out; the explanation is still unknown. Below the lambda point, helium called 'helium II' can flow without any apparent viscous resistance through fine capilaries or as a thin film over a surface; but its viscosity, as measured, for example, by the damping of an oscillating cylinder or disk in the liquid, remains finite. If a source of heat is supplied the liquid will flow towards it, especially through a porous plug. Conversely a flow caused mechanically produces heat transfer. One crude consequence is that of extremely high apparent thermal conductivity; another is the startling 'fountain phenomenon' (Fig. 3.22). The name *superfluidity* was given by Kapitza. There is an obvious analogy between the frictionless flow of electricity and that of He II; in addition, just as a heavy current by its magnetic field destroys super-conductivity, so there is a limit to the speed of the frictionless flow of the liquid.

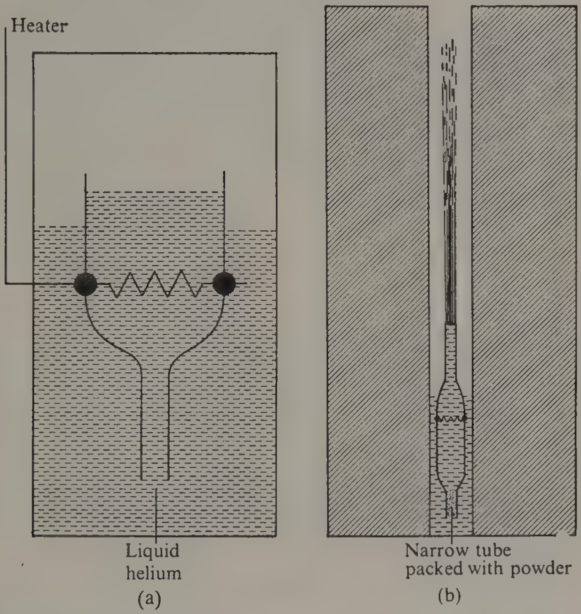


FIG. 3.22. Thermo-mechanical effect and helium fountain. When liquid helium in the cup is heated the level rises. The effect is greater the smaller the tube below the cup. By using many fine tubes in parallel the helium shoots out of the top of the cup like a fountain. From K. A. G. Mendelssohn, *Quest for absolute zero* (World University Library, 1961.)

The modern view is to regard both phenomena as in some way caused by the ordering of particles by velocity somewhat as atoms in a crystal are ordered in space, but no theory so far seems free from difficulties. The effects of the quantum become more noticeable as temperatures are lowered; for example, those on specific heats of solids (p. 58). Low temperatures thus reveal effects that, though the consequences of fundamental laws, are not seen at ordinary temperatures.

These two effects require for their detailed study temperatures much below the 1°K obtainable by evaporating liquid helium. These are reached by using the magnetism of certain salts, due to electronic spins. Magnetizing such a salt is analogous to compressing a gas; some of the energy goes into the magnetic field, but some into heat as in the gas. If this heat is first removed, demagnetizing the salt will cool it below its initial temperature, as compressed gas from a cylinder will cool if expanded into the air. By this method, repeated with different salts if necessary, temperatures of the order of 0·005°K can be made available for experiments.

Solid-state physics has benefited greatly from studies at low temperatures which are made possible by liquid helium; high magnetic fields are often associated with the low temperatures.

Such researches have been greatly helped by the easier access to liquid helium that has come from the commercial development of the Collins helium liquifier. This derives from Kapitza's machine in which helium is liquified by making it do mechanical work in an expansion engine.

Atomicity

The reader will have noticed that almost all the advances in physics we have described are not only the consequence of atomicity in nature but, unlike most of nineteenth-century physics, cannot be described without this idea. Under the word 'atomicity' I include all theories of which it is an essential part that certain quantities, if they exist at all, have integral values in terms of some natural units; thus it includes quantum theory as well as electron theory and the existence of nuclei with integral charges, each apparently composed of an integral number of nucleons.

When it is considered that up to the end of the last century some eminent chemists refused to accept the existence of atoms, preferring to speak only of 'combining weights', it is clear what a revolution has taken place. I believe this principle of integers to be the most fundamental fact yet discovered in physics and the only one we physicists have equal in generality with evolution in biology. In a modified form—the cell—nature has somehow managed to carry it over into biology; how is still unclear.

References

Aston, F. W. (1924) *Phil. Mag.* **47**, 385.
Becqueral, H. *C. r. hebd. Séanc. Acad. Sci., Paris* (1896) 122, pp. 420, 501, 689, 1086.
Bohr, N. (1913) *Phil. Mag.* ii 1–25, 476–502, 857–875.
Bragg, W. H. (1914) *Phil. Mag.* i 881–99.

Bragg, W. L. (1914) *Phil. Mag.* ii 355–60.
Broglie, L. de (1925) *Annls Phys.* **3**, 22.
Chadwick, J. (1932) *Proc. R. Soc.* A**136**, 692.
Cockcroft, J. D. and Walton, E. T. S. (1934) *Proc. R. Soc.* A**144**, 704.
Dirac, P. A. M. (1928) *Proc. R. Soc.* A**117**, 610 and **118**, 351.
Einstein, A. (1905) *Annln Phys.* **17**, 891.
Heisenberg, W. (1925) *Z. Phys.* **33**, 879.
Onnes, H. K. *Leiden Comm.* (1911) 199b, 120b, 122b: 147 and (1914) 142c: 126.
Planck, M. *Verh. dt. phys. Ges.* (1900) 237.
Rontgen, W. C. (1898) *Annln Phys.* **64**, 1.
Rutherford, E. (1919) *Phil. Mag.* ii 537–87.
Schrodinger, E. (1926) *Annln Phys.* **79**, 734.
Soddy, F. (1911) *The chemistry of the radioactive elements.* London.
Thomson, G. P. (1928) *Proc. R. Soc.* A**147**, 600.
Thomson, J. J. (1897) *Phil. Mag.*, 293.
Wilson, C. T. R. (1904) *Phil. Mag.* ii 681–90.

4 Geophysics

Introduction

As in other branches of science, the present century has seen an enormous development in the study of the physics of the Earth's interior. Geophysics has been, and still is, the meeting-place of a number of standard disciplines such as physics, chemistry, geology, astronomy, and dynamics, and the inferences that have been made concerning the interior of the Earth and the atmosphere above it have usually involved mathematical investigations, often complicated and lengthy. During the past fifty years the statistical treatment of data (long used by astronomers in presenting their results) has largely replaced the graphical methods often employed in earlier days, and the application of tests of significance is becoming customary.

For the solid Earth, in particular, direct sampling is practicable only in the outermost 10 km or so; the increase of temperature with depth rules out deeper drilling at present. Accordingly, information about the interior has to be obtained by inferences from many different kinds of observations, often involving a long chain of argument. To set out the subject in anything like its correct proportions necessitates a brief review of the legacy of the nineteenth and earlier centuries. Here, as throughout this chapter, the treatment will be illustrative rather than exhaustive.

The shape and size of the Earth can be determined by terrestrial measurements combined with astronomical observations; indeed, the radius was known to Eratosthenes, about 250 B.C. As a rough illustration, if the altitude of the pole-star is known at two points, one due north of the other, it is easy to see that this leads to an estimate of the radius R of the Earth. With refinements, this method will show that the 'figure' of the Earth is not that of a sphere but is nearly that of an oblate spheroid. The altitude of a star is specified with reference to the horizontal, as measured, say, with a spirit-level; the plane of the horizontal is thus at right angles to the direction of *apparent* gravity. Accordingly, the figure of the Earth is linked with the Earth's gravitational field.

The classical researches of Clairaut (1743) led to a formula concerning the mass, the principal moments of inertia, the flattening at the poles, and the ratio of the centrifugal force at the equator to mean gravity. Furthermore, this result is independent of any assumption about the distribution of mass within the Earth. The axis of the Earth does not point continually in the same direction, but describes a cone in space, of semi-angle about 23° (the 'obliquity of the ecliptic') in a period of 26,000 years (the 'precession of the equinoxes'). From Clairaut's formula, combined with the period of this precession, it was inferred that the moment of inertia C about the polar axis is about $0{\cdot}334\ MR^2$, where M is the mass of the Earth; M is found from Newton's theory of gravitation when the constant of gravitation is known, as for instance from the laboratory experiments of Boys. This value of C illustrates how the findings of different disciplines have been combined to give information about the interior of the Earth. Incidentally, since the moment of inertia is a measure of the central condensation it follows that the Earth cannot be uniform, for then the numerical coefficient would be 0·4. To anticipate more modern work, calculations show that it is very unlikely that the Earth is a chemically homogeneous body compressed under its own gravitation.

It will be clear that the preceding formula for C is not just an interesting piece of history. It remains a control on any suggested model of the internal constitution of the Earth. Further, Clairaut's theory led to a relation between gravity and latitude; the geodesist, in ascertaining the shape of the earth, therefore makes observations of gravity as well as measurements of position on the Earth's surface. Provided that geodetic measures can be connected by continuous survey, as in continental areas, the general form of the surface of an area can be computed from surveys; however, if the route of the survey reaches the sea, with no other land that can be reached by a radar or laser beam, the chain of observations is broken. It is, then, important to secure observations of gravity at sea.

Now this is not an easy task, for observations of gravity made in a ship contain the up-and-down accelerations in the motion of the ocean. In more recent times, it is of interest to record, a great advance was made by Vening Meinesz, who in 1923 made the first measures of gravity at sea by using a special pendulum apparatus during a long voyage in a submarine;

the observations could be taken at a sufficient depth for the effect of the motion of the sea to be greatly diminished. By observations at sea, combined with those on land, the gravitational field is becoming better known, but the extent of the oceans and continents is vast, and many years of work lie ahead. The attraction of irregularities in the outer layers of the Earth is likewise a complicating factor, though it has its compensations in guiding prospectors in search of oil and minerals.

The foregoing paragraphs illustrate the importance of earlier work on the terrestrial constitution and likewise the interrelation of different types of study. One other illustration may be given. The existence of volcanoes and the increase of temperature with depth in mines indicated great heat in the Earth's interior and led to the idea that the Earth has a molten interior, with a thin solid crust. It was Lord Kelvin who pointed out that in such circumstances the oceanic tides would be negligibly small; indeed, convergent lines of evidence, notably the propagation of waves from earthquakes and large explosions, have shown that the interior is rigid, on the average about as rigid as steel.

Modern trends

The present century has seen an enormous development in geophysical studies. Many factors have contributed to this, in addition to the general acceleration in the progress of science. On the economic side the need for accurate weather forecasting has brought much-increased attention to theoretical meteorology; the search for oil, gas, and minerals has emphasized the importance of the various methods of geophysical prospecting; polar exploration has drawn attention to a whole range of new problems; the possibilities of food production and the influence on climates have greatly stimulated the science of oceanography; the construction of earthquake-proof buildings and other structures, particularly in developing countries, is linked with theoretical seismology. Military necessity and the interest in travel in space have played a considerable part, and the use of high-speed electronic computers has made feasible the solution of many problems for which previously computation by existing methods had been far too slow a process of research.

It will be convenient to treat geophysics as a collection of separate disciplines. This is perhaps inevitable, and it is the manner in which international and national geophysical studies are organized. However, as we have already seen, the Earth presents its problems in its own way, and an indication of the way in which these disciplines are interrelated is seen in the increasing number of symposia in which geologists, geophysicists, and geochemists are pooling the results of their researches. The fundamental drive in attempts to widen our knowledge of the interior of the Earth and of the regions just outside it is undoubtedly human curiosity. As with other branches of science, the solution of a problem leads to other researches in its train. Our elation at new developments is tempered by our realization of our ignorance in the face of the growing number of problems awaiting investigation.

An enormous literature exists on the subject of the origin of the solar system, and of the Earth in particular; however, the only general statement that it is safe to make is that the subject is being widely studied but is still in a highly controversial state.

Seismology

Our most detailed knowledge of the interior of the Earth has been derived from a study of earthquake records. The method is to interpret so far as practicable the motion of the ground at the recording station. For this purpose a number of instruments had been devised during the nineteenth century, and mention must specially be made of the work of Milne and his collaborators during the twenty years that they spent in Tokyo from 1876 to 1895 at the invitation of the Emperor of Japan; this work led to the setting-up of a flourishing school of earthquake research in Japan. Milne continued his researches on his return to England in 1895, and the Milne seismograph (of the ‘horizontal pendulum’ type), greatly improved by Shaw with the introduction of optical recording on photographic paper, brought increased sensitiveness; until comparatively modern times the Milne–Shaw instrument was operated at a network of observatories installed at stations in the British Empire. The effort has continued to improve the sensitivity, and a most important step forwards was the construction by Prince Galitzin in 1909 of instruments incorporating electromagnetic recording, in which a coil fixed to the moving arm of the seismograph moves between the poles of a fixed magnet; this device is used in most modern instruments.

Modern seismology may be considered to date from the recognition by a geologist, Oldham, in 1897 that the two relatively small onsets seen at the beginning of a seismogram would correspond to the propagation through the Earth’s interior of compressional and distortional waves, in agreement with the theory of Poisson (1829). These two onsets were later named *P* and *S* respectively—an allusion to their earlier name as ‘primary and secondary phases of the preliminary tremors’—or ‘*undae primae*’ and ‘*undae secondae*’. A further important stage was the identification by Oldham of the large waves, constituting the principal part of the records, as ‘Rayleigh waves’, predicted theoretically by Lord Rayleigh in 1887. A difficulty later emerged in that the motion in waves of Rayleigh’s type should be wholly in the plane of propagation, whereas motion transverse to this plane existed; the matter was cleared up by Love in 1911, who showed that waves with purely transverse motion could exist if a surface layer were underlain by material in which the velocity of distortional waves was greater than in the layer. Such waves are now called ‘Love waves’. He also showed that in a medium in which this velocity increases with depth the motion in Rayleigh waves, like that of Love waves, is dispersive, i.e. the velocity depends on the period.

Body waves

For any given earthquake it became apparent, from the times of arrival of *P* and *S* at different stations, that to a first approximation the Earth could be regarded as spherically symmetrical; with increasingly accurate timing (by putting time signals on the record) it was possible to draw up a table of times of travel of *P* and *S* to different angular distances round the world. These times were uncertain on account of the differing depths of the origin (the ‘focus’) of the earthquakes studied, but were good enough to make possible a rough location of a shock even though no observations were available from observers on the spot. From the study of a large number of well-observed earthquakes improved tables were drawn up, and these in turn led to a more precise location of the place (the ‘epicentre’) below which the shock occurred. Such tables are the basis of much of the information now available about the properties of the Earth’s interior. This process of location is nowadays conducted on an international basis.

Now, just as in optics, a table of times of travel can be analysed to give the form of the earthquake rays and the velocity of *P* or *S* at all depths reached. This is still a main line of advance. It was early noted that at an angular distance $\Delta = 105°$ from the epicentre, the *P*-onset changed its character, while *S*, as such, ceased to be recorded. The interpretation of this phenomenon by Oldham (1906) was that the Earth possesses a ‘core’, which, unlike the outer ‘shell’ or ‘mantle’, does not transmit distortional waves, and this is the characteristic of a fluid. The depth of the boundary of the core was calculated by Gutenberg in 1914 as 2900 km, a figure supported by later calculations. From Gutenberg's work a rapid development ensued. He computed the approximate times of travel of a number of different waves that had undergone reflexion, refraction, and transformation; these were duly identified in lists of observed onsets or on seismograms, and by about 1930 a fairly good idea of the elastic properties of the Earth's interior was available.

The time was now ripe for further advances. Jeffreys and Bullen in 1935 published tables of times of *P* and *S* based on a large number of well-observed shocks, and these tables, refined in 1940, constitute the '*J.-B. Tables*' still in use in reducing observations. The question of the internal variation of density was in the main worked out by Bullen in 1936, and by combining the density and the velocity of *P* and *S* waves the elastic constants were determined by Bullen for all depths reached by these waves. An extremely important discovery was made in 1936 by Lehmann; observations of the supposed 'diffracted P' beyond $\Delta = 105°$ led her to the idea that an 'inner core' existed. Calculations by Gutenberg and Richter and also by Jeffreys in 1938 gave the radius of the core as about one-fifth of the radius of the Earth's surface.

Observations made at stations within a few hundred kilometres of an earthquake are specially valuable, and have been used by various workers, notably Mohorovičić (1909), Conrad (1925), and Jeffreys (1927) to delineate the crustal layering of the continents. Artificial explosions, including nuclear explosions, have given similar information; the structure of the sea-floor has been widely examined in this way, which also is often used in seismic prospecting, as mentioned earlier.

Surface waves

The foregoing use of body waves in elucidating continental and oceanic structure finds its complement in studies of the passage of Rayleigh waves and Love waves. Here the method is to assume a model structure and work out the relation between wave period and velocity. Technically the velocity read from seismograms is the 'group velocity' and not the wave velocity. By comparison, the likely *average* constitution of a region of continental dimensions is arrived at.

Seismicity and earthquake magnitudes

The term 'seismicity' is used for the distribution of earthquakes and their intensity. The distribution has long been known in general terms, and lists of epicentres compiled from the routine computations already referred to enables belts of seismic activity to be delineated. An important advance, putting the subject on a quantitative basis, was made by Richter in 1935 by the introduction of a scale of earthquake magnitudes derived from readings of seismographs; the magnitude is related to the energy released by the shock.

It is thus possible to compute the rate at which energy is being released in earthquake shocks. This release of accumulated strain promises to be one of the indications of large earthquakes and offers some hope of predicting such disasters.

Modern advances in recording

The fundamental idea in constructing a seismograph has been, until relatively recent times, to have a mass or pendulum bob so suspended or pivoted that it is set in motion relative to the base of the instrument by any movement of the Earth. In 1935 a new principle was introduced by Benioff; the 'strain seismograph' records the distortion of the Earth in a given direction by measuring electromagnetically the variation in the distance of two points of the ground. These points may be of the order of 20 metres apart, and this form of seismograph is particularly valuable in recording earthquake waves of very long period, even up to 1 hour.

One of the troubles with most recording instruments is the occurrence of unwanted signals, usually more or less random in character. The aim of the experimenter is usually to arrange for the signal to show up against the 'background noise'. Sometimes the general frequency of the background noise is such that it can be 'filtered out'—usually electrically. Even when no recognizable earthquake appears on a seismogram the trace is not at rest, but shows a microseismic movement, in part fairly regular, in part random. In recent years it has proved possible to use an 'array' of seismographs to increase greatly the signal-to-noise ratio. For instance, the array at Eskdalemuir in Scotland has twenty-one seismometers arranged in two arms at right angles forming a cross; the distance between consecutive seismometers is 980 yards. The outputs are recorded on magnetic tape, and combined in suitable ways so as to allow the 'noise' largely to cancel out. The high precision of these instruments is useful in detecting distant explosions as well as distant small earthquakes.

The oscillations of the earth

A number of calculations have been made, notably by Pekeris (1937) and his collaborators, of the periods of oscillation to be expected in the Earth, using accepted distributions of density and elasticity. A dramatic verification of this work was the announcement in consecutive contributions to a meeting of the International Association of Seismology and the Physics of the Earth's Interior (Helsinki, 1960) that free periods of oscillation, agreeing with the relevant calculated periods, had been found in the analysis of the observations of the Chilean earthquake of 22 May 1960; Benioff, Press, and Smith (1961) analysed seismograph records, and Ness, Harrison, and Slichter (1961) examined the records of an extremely sensitive gravity meter. The close accordance with calculation confirms at the same time the reality of the modes of oscillation and the accuracy of the Earth model used in the computations. Later work has abundantly verified these findings.

Terrestrial magnetism

The Earth's magnetic field, measured at any one place, is continually changing. These variations have been measured for many years at a number of magnetic observatories; they are partly periodic, with astronomical periods, partly continuous, and partly sporadic, particularly as 'magnetic storms', which are associated with outbursts of activity in the Sun.

The general magnetic field roughly corresponds to that of a magnet of small size at the centre of the Earth, or equivalently to that of a uniformly magnetized sphere. The origin of this 'dipole field', the strength of which is slowly changing, has been the subject of much controversy, and as yet there is no general agreement on the subject. The theory that has found favour in recent years is that of a self-exciting dynamo, as originally suggested by Larmor in 1919: the possibility of the existence of such a dynamo was established by Backus (1958) and Herzenberg (1958). In 1949 Bullard had given a discussion of a definite model, and with Gellman in 1954 had obtained approximate solutions of the basic equation of a steady dynamo.

Superposed on this dipole field is a 'non-dipole field'; at any place on the Earth this is slowly changing its direction, and this 'secular change' has been measured over 300 to 400 years at a number of observatories. Many geographical maps, for instance, show the magnetic north, with a note of the annual variation. In general terms, the magnetic north pole of the Earth is moving around the geographic north pole at a rate of about one revolution in 500 years. The non-dipole field also shows secular changes.

An analysis by Lahiri and Price in 1938 indicated that both for periodic variations and for storm-time variations the electrical conductivity of the Earth increases greatly with depth in the neighbourhood of a depth of 600 km.

Magnetic maps constructed for the whole surface of the Earth show, both for the vertical component of the magnetic field and for the rate of change of the field, a pattern with the distinctive feature that the contours of equal vertical force or of equal rate of change form a small number of families of oval curves, the middle points of which correspond to a maximum or minimum value. The area covered by any one family is of continental size. This pattern shows a general westward drift. This westward drift, which is well-established, has been attributed to the movement of eddies within the Earth's core; the interaction between an eddy and the internal magnetic field would be equivalent to the existence of a magnetic dipole.

Palaeomagnetism

Most substances acquire some magnetization when placed in a magnetic field. For some bodies, such as sulphur, the amount of this induced magnetization is small, and disappears when the magnetic field is removed. For some other substances, such as steel, the induced magnetism is large, and may be permanent. Such substances are called 'ferromagnetic'; many rocks behave in this manner, owing to the presence of inclusions of ferromagnetic minerals. Ferromagnetic materials, magnetized in the laboratory, acquire a 'remanent magnetization', depending on the magnetic field applied. The natural remanent magnetization of rocks is thought to bear the same relation to the field in which magnetization occurred. Since the magnetization disappears when the substance is heated to the 'Curie point', the remanent magnetization may be expected to indicate the direction and intensity of the Earth's magnetic field during the period of cooling. Other factors enter into the discussion, however, and the line of argument is often complicated. For instance, sediments consisting partly of magnetic and partly of non-magnetic particles acquire a remanent magnetization during deposition, as is shown by experiments in the laboratory; this magnetization may not be exactly parallel to the applied field, and further, its direction may be affected by currents in the water. If the sedimentary rocks are subsequently folded, as often happens, due allowance must be made for this folding in interpreting the directions of magnetization in rock specimens.

The phenomena of remanent magnetization are more complex than the foregoing outline suggests. Considerable interest has been shown in these phenomena in recent years, and an enormous number of studies have been made, notably under the stimulus and direction of Runcorn, of the direction of remanent magnetization in rock specimens of known orientation. The mode of analysing these measurements is too complicated to discuss here. A convenient way of looking at the measurements is to find the position the magnetic pole would have if the Earth's field were a true dipole field. Each observation of direction of magnetization would then give a 'virtual geomagnetic pole' (V.G.P.). Now for a group of rocks of the same age and from the same general locality the V.G.P.s often fall into two groups, 180° apart. Thus, for rocks of ages up to some 30 million years about one-half have a 'reversed magnetization'; the view is widely, but not unanimously, held that the Earth's field itself underwent reversal.

The positions of the V.G.P. for specimens of any given geological period cluster fairly closely round a direction that may be taken as the V.G.P. for that period. Maps have accordingly been produced showing the way in which the direction of the Earth's magnetic axis has moved, relative to the continents, through geologic time, from the Pre-Cambrian (over 600 million years ago) to the present time. As an indication of the magnitude of these changes, it may be mentioned that in Carboniferous times the geomagnetic pole was somewhere in the general neighbourhood of Japan instead of at its present position in north Greenland. The changes in historic times have been found from measures from clay in well-dated sites, such as fireplaces or kilns. Much discussion has taken place over the significance of this polar wandering: the Earth's field may not always have been that of a dipole; the Earth's axis of rotation may have wandered with respect to the crust; large portions of the Earth's surface may have moved relative to one another; the Earth may have expanded. So much research still needs to be done that it is premature to try to give a final interpretation of these apparent changes.

Artificial satellites

A satellite describes an orbit under the gravitational field; all the other heavenly bodies make their contribution to this field, but for the most part the field arises from the attraction of the masses within the Earth. We have already seen that astronomical observations lead to a knowledge of the moment of inertia of the Earth about its axis. Although the Earth is on the whole rigid for forces of short period, such as those involved in the passage of earthquake waves, it has, presumably over long periods of time, assumed approximately the hydrostatic state, that is, the shape it would have if it were fluid.

Now if a satellite orbit is observed with high accuracy for a sufficient time the gravitational field can be deduced with precision. In recent years it has been shown that this field departs measurably from what would be its value for a fluid Earth. Thus, observations of satellites indicate that the Earth is not exactly in a hydrostatic state, and the departure from this state can be estimated the more exactly as observations of the orbits of satellites increase in number and accuracy.

Evidence from the tides

The tide-raising force can be calculated from the attraction of the Sun and Moon at any place; the theory goes back to Newton. In relation to the Earth, the Sun and Moon move in nearly elliptic orbits, which are not in the plane of the Earth's equator. This force is conveniently decomposed into a large number of terms, with periods corresponding to a lunar half-day, a solar half-day, a sidereal half-day, and a number of others, together with daily, fortnightly, monthly, and semi-annual periods, amongst others. Of these, the lunar semi-diurnal force is the most important. Each of these terms gives rise to an oceanic tide of corresponding period, and these can be separated at any place, such as a harbour, where a tide-gauge is installed.

Because the Earth is not completely rigid it yields to some extent to the tidal forces, giving 'Earth tides'. These can be measured in the records of a sensitive gravimeter or of a seismograph of horizontal pendulum type. The amplitude of these tides gives a measure of the overall elasticity of the Earth; the Earth tide affects the observed height of the oceanic tide, partly because of the corresponding rise and fall of the ocean floor, and partly because of the attraction of the 'tidal bulge' in the Earth and the ocean. The study of Earth tides is of great geophysical importance, and in recent years has received much-increased attention.

It will be recalled that the existence of oceanic tides was cited by Lord Kelvin as evidence that the Earth is not a fluid with just a thin shell. The theory of the tides is very complicated, particularly as the problem is a dynamical one and not a statical one. It would scarcely be feasible to summarize here the many researches that have been carried out in modern times.

These same forces must raise tides in the atmosphere, and early last century Laplace gave considerable attention to the problem. The search for a half-daily lunar tide (the tide most likely to be detected) is a fascinating story, bringing the subject up to modern times, and is worth setting out in some detail. It is a beautiful example of the way geophysics is the meeting-place of various disciplines.

Tidal movements in the atmosphere should cause corresponding variations of barometric pressure at ground level. The behaviour of the barometer is very different in the tropics from what is found in temperate latitudes. In the tropics barometric records clearly show an oscillation with period one-half a *solar* day, i.e. these oscillations are on the whole due to the Sun and are not linked with *lunar* time, but may, of course, have a lunar term superposed: in fact, the analysis of 17 months of 2-hourly readings of the barometer at St. Helena enabled Captain Lefroy to detect in 1842 the air tide due to the Moon, a discovery supported by many later determinations.

In temperate latitudes the story is very different. The large changes of the height of the barometer in a few hours are familiar to everybody living in western Europe. Looking backward, it is thus not so surprising that a number of attempts last century to determine the air tide were unsuccessful. For instance, Sir George Airy, the Astronomer Royal, analysed without success some 180 000 hourly observations of the barometric height at Greenwich for the years 1854–73, and he announced in 1877: ‘We can assert positively that there is no trace of lunar tide in the atmosphere’. This assertion was not disproved until 1918, when Chapman adopted a new attack on the Greenwich data, which by then amounted to 64 years of observations; Chapman’s procedure was to reject readings for all the days on which the range of pressure was less than one-tenth of an inch, even though this meant rejecting two-thirds of the readings. In this way the amplitude of the lunar half-day tide at Greenwich was found to correspond to a pressure of 0·01 mm of mercury. Parallel investigations of the air tide have since been made, many by Professor Chapman and his co-workers, so that the lunar air tide is now well-determined (Chapman 1935).

Other aspects, however, remain to be considered. The periodic compressions and rarifications should give rise to a semi-diurnal fluctuation of temperature; this was confirmed by Chapman, who from observations of the temperature at Batavia (now Jakarta) from 1866 to 1928 found an amplitude of a little less than 0·01°C. Further, he pointed out that the tidal movement of the air at high levels in the Earth's magnetic field should give rise to electric currents, which in turn should give a lunar half-daily variation in the Earth's magnetic field. This he duly disentangled from a long series of observations at magnetic observatories. We must further mention the lunar variation in the height of the *E*-layer; this was detected by Appleton and Weekes in radio measurements at Cambridge in 1939, but the amplitude, about 1 km, is far greater than had been expected.

The *solar* semi-diurnal tide in the atmosphere was for long a puzzle; it is obvious in a record made in the tropics, while the corresponding lunar tide has to be analysed out from a long series of observations. To account for this large amplitude, Lord Kelvin in 1882 suggested that it might be a resonance phenomenon, such as would happen if the atmosphere had a natural period of oscillation near to 12 hours. However, the times of travel of air waves from the great Krakatoa explosion of 1883 to different meteorological stations suggested a period of about 10½ hours. The theoretical problem of the free oscillations was tackled in 1937 by Pekeris, who used all the data then available for the variation of temperature with height. He found the rather unexpected result that the atmosphere possesses, in addition to higher modes, oscillations of approximately both these free periods, 12 hours and 10½ hours. The amplitude of the 12-hour oscillation at a height of 100 km would be about 200 times that at the ground, thus accounting for the large amplitude of the oscillations in the *E*-layer. Later work by Wilkes substantially confirmed these results. Their importance at the time was the information they gave concerning the variation of temperature at great heights, particularly from 80 to 100 km. It is very satisfactory to know that these results have since been verified by measurements made during rocket ascents.

The earth's thermal history

The temperatures of the surfaces of the planets, and of the Earth in particular, are determined by the balancing of the heat received continually from the Sun and heat radiated from the surface into space: this mechanism is confirmed by astronomical measurements of the surface temperatures of the planets. In addition to this approximate balancing there is a continual flow outwards from the heat stored in the Earth at any one period. This heat flow is known at a large number of places on land and at sea, from measures of the temperature gradient in borings and in ocean sediments. It amounts to about one-millionth of a calorie per second per square centimetre of the surface, and this is about the same over land as over the ocean floor; as compared with the mean rate at which heat is received from the Sun, somewhat less than one-hundredth of a calorie per second per square centimetre, this is a ratio of the order of only one ten-thousandth. Nevertheless, the heat outflow is an important geophysical datum.

Some of the heat flowing out may represent a running-down of the store of primitive heat remaining from the time of formation or of solidification (if, indeed, the Earth was at one time molten). The remainder is presumably the heat generated by radioactive disintegration of uranium, thorium, and potassium. The evidence of laboratory measurements is that acidic rocks, such as the granites, generate more heat than basic rocks (e.g. basalt), and that these in turn are more radioactive than the ultrabasic rocks that are thought to comprise the material down to the Earth's core.

The composition of the continents is known in general terms from seismological studies, and in the main it is possible to account for the observed heat flow from the surface, using the probable content of radioactive materials in the crust, the heat generated during disintegration, and the known thermal conductivities of the rocks of the crust and the mantle. The correspondence is fairly satisfactory, but there is an outstanding difficulty. Earthquakes and explosions indicate a continental crust, with basic rocks below a granite layer, of total thickness about 33 km, whereas underlying the thin layer of sediments covering the ocean floor to a depth of 2 to 4 km there appears to be a basaltic layer only about 4 km thick, which would not account for the heat flow corresponding to the continents. The material of the mantle is to some extent radioactive, but to make the heat flow over the oceans compatible with that over the continents suggests that the composition of the mantle below the oceans differs from what is found for continental regions. The difference has not yet been satisfactorily explained without introducing *ad hoc* hypotheses,

It seems then that the problem of heat flow may be linked with the composition of the sub-continental rocks. This, in turn, brings in the question of the origin of the Earth, and indeed of the solar system. If we knew the distribution of temperature at some specific time in the past, as well as the contribution from sources of heat since that time, then the theory of heat conduction would lead to a unique solution. A number of calculations have been made along these lines, notably by Jeffreys, but the initial distribution of temperature has to be derived from some hypothesis such as the solidification from a molten state. To discriminate is not easy; heat conduction is a slow process, and temperature differences tend to become smoothed out, so that different hypotheses can lead to much the same value of the temperature and rate of heat flow at the surface, where, indeed our observations have to be made.

The age of the Earth

The expression 'age of the Earth' is not a very precise one; for the present purpose it may be taken as the time that has elapsed since a firm crust was formed. This initial time is near enough for most purposes, in view of our ignorance of the history of the primitive Earth; if the Earth was fluid at one time the cooling to form a thin crust would be a relatively brief process.

Various methods have been proposed for ascertaining the age of the Earth—most of them, like the rate of accumulation of salt in the sea, subject to large uncertainties. Astronomical considerations suggest five or six thousand million years. The most reliable indication is that given by the age of the oldest rocks, which presumably are younger than the solid crust of the Earth. The age of these rocks can be found from the uranium/lead ratio on the assumption that this depends on radioactive disintegration; the oldest known rocks are about four thousand million years old. Considering the rapid development of cosmology in modern times this approximate agreement is satisfactory, but is scarcely acceptable as a final determination.

Conclusion

In this chapter an attempt has been made to indicate how different branches of knowledge and different lines of attack have been brought together in order to throw light on the composition and state of the Earth's inaccessible interior; this has not been the place to present a catalogue of results, and the presentation necessarily has to be given in outline form. Precision may be limited by the inaccuracies of observational data, but in deciding among several hypotheses an argument that gives the order of magnitude is often a valuable guide, and sometimes indeed the only one. As Sir Harold Jeffreys has pointed out, incorrect hypotheses often fail by large margins.

It has not been practicable to discuss here the findings of modern geochemistry, important though they are. 'Phase changes', to which it has been suggested the discontinuities within the body of the Earth may correspond, bring in the questions of lattice structure and atomic physics. The recognition of the many outstanding difficulties of the subject is indeed an indication of the healthy development of the science of geophysics.

Selected References

Appleton, E. V. and Weekes, K. (1939) On lunar tides in the upper atmosphere. *Proc. R. Soc.* A**171**, 171–87.

Backus, G. E. (1958) A class of self-sustaining dissipative spherical dynamos. *Ann. Phys.* **4**, 372–447.

Benioff, H., Press, F., and Smith, S. (1961). Comptes Rendus. International Association of Seismology and the Physics of the Earth's Interior (Helsinki, 1960).

Bullard, E. C. Electromagnetic induction in a rotating sphere. (1949) *Proc. R. Soc.* A**199**, 413–34.

—— and Gellman, H. (1954) Homogeneous dynamos and terrestrial magnetism. *Phil. Trans. R. Soc.* A**247**, 213–78.

Bullen, K. E. (1936) The variation of density and the ellipticities of strata of equal density within the Earth. *Mon. Not. R. astr. Soc. geophys. Suppl.* **3**, 395–401.

Chapman, S. (1918) An example of the determination of a minute periodic variation, as illustrative of the law of errors. *Mon. Not. R. astr. Soc.* **78**, 635–8.

—— The lunar tide in the Earth's atmosphere. (1935) *Proc. R. Soc.* A**151**, 105–17.

Conrad, V. (1925). Laufzeitkurven des Tauernbeben von 28 November 1923, *Mitt. Erdb-Kommn Wien* **59**.

Gutenberg, B. (1914) *Nachr. Ges. Wiss. Göttingen* **1914**, 1–52, 125–76..

Herzenberg, A. (1958) Geomagnetic dynamos. *Phil. Trans. R. Soc.* A**250**, 543–85.

Jeffreys, H. (1926) On near earthquakes. *Mon. Not. R. astr. Soc. geophys. Suppl.* **1**, 385–402.

—— (1927) *Beitr. Geophys.* 1–29.

—— (1959) *The Earth: its origin, history, and physical constitution*, 4th edn. Cambridge University Press.

—— and Bullen, K. E. (1940) *Seismological tables*. British Association, London.

Lahiri, B. H. and Price, A. T. (1938) Electromagnetic induction in non-uniform conditions, and the determination of the conductivity of the Earth from terrestrial magnetic variations. *Phil. Trans. R. Soc.* A**237**, 509–40.

Larmor, J. (1919) How could a rotating body such as the sun become a manget? *Rep. Br. Ass. Advmt Sci.* **1919**, 159–60.

Lehmann, I. (1936) *Publs Bur. cent. sism. int. U.G.G.I.* A**14**, 87–115.

Love, A. E. H. (1911) *Some problems of geodynamics*. Cambridge University Press.

Meinesz, F. A. Vening (1948). Gravity measurements as sea 1923–38. *Publ. Neth. Geod. Comm.* Delftsche Uitgevers.

Mohorovičić, A. (1910) Das Beben vom 8 X 1909. *Godiš. Izv. Kr. zem. Zav. Met. Geodin.* **9**, pt. 4, 1–63.

Ness, N. F. Harrison, J. C., and Slichter, L. B. (1961) Observations of the free oscillations of the Earth. *J. geophys. Res.* **66**, 621–29.

OLDHAM, R. D. (1906) On the constitution of the interior of the Earth, as revealed by earthquakes. *Q. Jl geol. Soc. Lond.* **62**, 456–73.

PEKERIS, C. L. (1937) Atmospheric oscillations. *Proc. R. Soc.* A**158**, 650–71.

RICHTER, C. F. (1935) An instrumental earthquake magnitude scale. *Bull. seism. Soc. Am.* **25**, 1–32.

RUNCORN, S. K. (1962) Palaeomagnetic evidence for continental drift and its geophysical cause. In *Continental drift* (edited by S. K. Runcorn). pp. 1–39. Academic Press, New York.

General References

COOK, A. H. and GASKELL, T. F. (editors). (1961) The Earth today. *Geophys. J.* **4**.

GASKELL, T. F. (1967) *The Earth's mantle*. Academic Press, New York.

IRVING, E. (1964) *Palaeomagnetism and its application to geological and geophysical problems*. Wiley, New York.

KUIPER, G. P. (editor) (1954). *The Earth as a planet. The solar system*, Vol. 2. University of Chicago Press.

5 *Chemistry*

BY the middle of the nineteenth century the distinction between chemical atoms and chemical molecules had been established, particularly by the work of Avogadro and Cannizzaro. Their theory of matter served to account for the principles that had been discovered to govern the combination of elements into compounds, by such chemists as Dalton, Berzelius, Gay-Lussac, and Wollaston. The problem of the means by which atoms combined to form molecules came to occupy the attention of chemists more and more. To understand the significance of the developments of theoretical chemistry in the twentieth century, when techniques for the solution of this problem were developed, it is necessary to look closely at the developments that took place in theoretical chemistry, particularly in the theory of valency and the chemical bond, in the forty years that followed the publication of Cannizzaro's *Memoir* in 1859.

'When the formulae of inorganic chemical compounds are considered', wrote Frankland, in 1852, 'even a superficial observer is struck with the general symmetry of their construction; the compounds of nitrogen, phosphorus, antimony and arsenic especially exhibit the tendency of these elements to form compounds containing three or five atoms of other elements [such as chlorine]. . . . Without offering any hypothesis regarding the cause of this symmetrical grouping of atoms, it is sufficiently evident . . . that such a tendency or law prevails, and that . . . the combining power of the attracting element, if I may be allowed the term, is always satisfied by the same number of these atoms.'

Because molecules of the simplest compounds between hydrogen or chlorine and another element often contain several atoms of hydrogen or chlorine but never more than one atom of the other element, the 'combining power', 'atomicity', or 'valence' of hydrogen and chlorine was set at one, making that of oxygen and sulphur (in, for example, H_2O and H_2S) two, and that of nitrogen (in, for example, NH_3) three. By far the most important application of this conceptual scheme for predicting or interpreting, Dalton's 'small whole numbers' was to be the element carbon.

'If only the simplest compounds of carbon are considered', wrote Kekulé (b. 1829) in 1858, referring to methane (CH_4), carbon tetrachloride (CCl_4), methyl chloride (CH_3Cl), chloroform ($CHCl_3$), phosgene ($COCl_2$), carbon dioxide (CO_2), carbon disulphide (CS_2), and hydrogen cyanide (HCN), 'it is striking that the amount of carbon which the chemist has known as the smallest possible, as the atom, always combines with four atoms of a monatomic element, or two atoms of a diatomic element; that in general the sum of the chemical equivalents of the elements which are bound to one atom of carbon is equal to four. This leads to the view that carbon is tetratomic (or tetrabasic)', or, we would say today, 'tetravalent'.

Immediately, however, there arises the question: Is carbon tetravalent in a compound such as C_2H_6? In such cases, wrote Kekulé, 'One must assume that some at least of the atoms [here hydrogen] are held in the compound in the same way [as in methane, for example] by the affinity of the carbon, and that the carbon atoms themselves align themselves next to each other, whereby a portion of the affinity of the one carbon atom is, of course, bound by an equal portion of the affinity of the other.'

Thus, just as Avogadro saved his idea that equal volumes of gases contain equal numbers of molecules by introducing a new concept, the polyatomic molecule of an element, so Kekulé rescued his idea that carbon is tetravalent by inventing the carbon–carbon bond. Fuller justification for the assumption that 'carbon enters into chemical union with itself'—an assumption that is the backbone of structural organic chemistry—was given in the same year, 1858, by a young Scottish chemist, Couper (b. 1831).

'If the four atoms of hydrogen [in CH_4] were bound together', observed Couper, then, in view of the existence of CCl_4, and by the very nature of the concept valence, 'we could evidently expect to form such bodies as H_4Cl_4, H_4Cl_3Br, $H_4Cl_2Br_2$, etc. . . . These bodies are not only unknown, but the whole history of hydrogen might be investigated and not a single instance be found to favour the opinion that it has any affinity *for itself* [italic added] when in union with another element. On the other hand, carbon remains chemically united to carbon, while perhaps 8 equivalents of hydrogen are exchanged for 8 equivalents of chlorine, as in naphthaline. . . . All the countless instances of substitution of chlorine, etc. [for hydrogen] tend in the same direction. They prove beyond doubt that carbon enters into chemical union with carbon, and that in the most stable manner.'

Neither Kekulé nor Couper discussed explicitly, in 1858, the nature of carbon–carbon linkages in such carbon-rich 'combinates' as C_2H_4 and C_2H_2, although Kekulé did remark that for them 'a denser arrangement of carbons must be assumed'. How the denser arrangement of carbons in such 'unsaturated' compounds might be represented, and the tetravalence of carbon preserved, was soon suggested by Crum Brown (b. 1838), a colleague of Couper's at Edinburgh, and by Erlenmeyer (b. 1825).

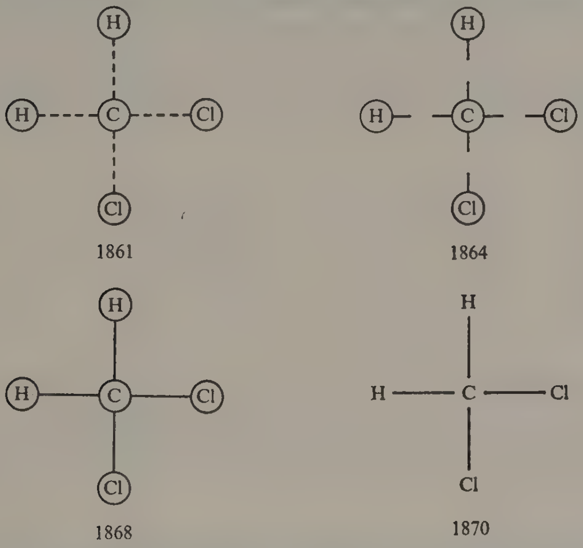


FIG. 5.1. The evolution of graphic formulas. Crum Brown's representation of CH_2Cl_2 (dichloromethane).

In effect, Crum Brown invented the first working model of the chemical bond. The evolution of his representation of this concept, illustrated for the molecule CH_2Cl_2 (dichloromethane), is shown in Fig. 5.1. In 1864, to represent ethylene (C_2H_4), Crum Brown invented the double bond; two years earlier, to represent acetylene (C_2H_2), Erlenmeyer had introduced the triple bond (Fig. 5.2).

These 'graphic formulae' or 'bond diagrams', though comparable in importance (in chemistry) to the symbols of the elements, are still a language 'not understood of the people', nor yet, perhaps, fully appreciated by chemists, although for nearly a century they have been the chief medium of expression and transmission of chemical knowledge in organic chemistry. Since Crum Brown's time they have received one important refinement.

If Crum Brown's planar formulae are taken seriously, one would expect to find two substances with the formula CH_2Cl_2, one in which the chlorine atoms of each molecule are adjacent to each other, as shown in Fig. 5.1, and one in which the chlorine atoms are opposite each other.

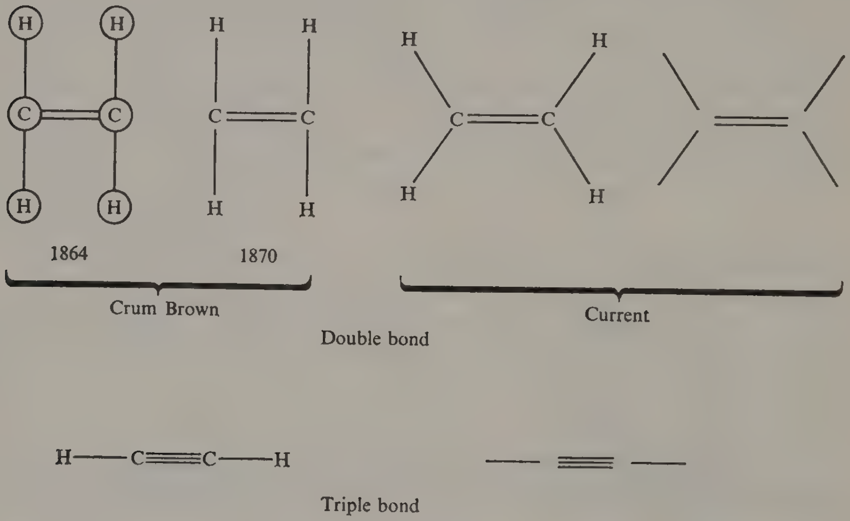


FIG. 5.2. Double and triple bonds. Crum Brown's graphic formulae for C_2H_4 (ethylene), together with the modern graphic formulae for C_2H_4 and C_2H_2 (acetylene). In these graphic formulae, each symbol 'C' is surrounded by four valence-strokes.

In fact, only one substance with the formula CH_2Cl_2 has been synthesized. This result was interpreted to mean that the four valencies of a carbon atom are directed not in a plane, as shown in Fig. 5.1, but towards the corners of a tetrahedron, at the centre of which lies the carbon atom, as shown in Fig. 5.3 for the molecule CH_3Cl. In the tetrahedral model, though not in the corresponding planar model, of CH_3Cl, all three hydrogen atoms are equivalent. Thus, it makes no difference which one is substituted for in going to the dichloro-derivative: there will be only one product with the formula CH_2Cl_2.

For ten years chemists did not take seriously the tetrahedral model of the carbon atom. Then, in 1874, a young student, van't Hoff (b. 1852), still two years away from a minor academic post at a veterinary college in Utrecht, legalized the tetrahedral model by showing that it was useful in interpreting a *physical* effect, the rotation of the plane of polarized light. Further, van't Hoff showed how, through the sharing of corners, edges, and faces, the tetrahedral model could be used to account for the occurence of single, double, and triple bonds (Fig. 5.4).

So stood the theory of carbon compounds, nearly complete in itself, but apparently wholly unrelated to other areas of chemistry, and seemingly completely contradictory to the basic premises of Newtonian physics. How, for example, could the attractive forces emanating from a carbon atom be independent of the positions of the attracted bodies, as van't Hoff seemed to postulate? Worse yet, if the attractive forces of a carbon atom do emanate in tetrahedral directions, then in unsaturated compounds, such as C_2H_4 and C_2H_2, they do not even point toward the attracted bodies. Little wonder that, in 1891, Hermann von Helmholtz (b. 1821), master of thermodynamics, electrodynamics, and hydrodynamics, wrote, 'The whole extraordinarily comprehensive system of organic chemistry has developed in the most irrational manner, always linked with sensory images which could not possibly be legitimate in the form in which they are presented.' Still, it was Newton himself who had said, 'We must learn from the Phaenomena of Nature what Bodies attract one another, and what are the Laws and Properties of Attractions, before we enquire the Cause by which the Attractions is perform'd.'

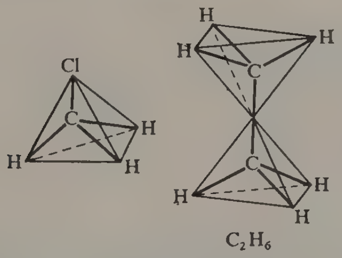


FIG. 5.3. Tetrahedral model of CH_3Cl. Note that the three hydrogen atoms are equivalent; i.e. the model is turned into itself by one-third of a revolution about the carbon-chlorine bond.

In retrospect, it seems unlikely that, solely from a study of the chemical transformations of matter, in test-tubes, chemists would ever have discovered the Cause by which chemical attraction is perform'd, any more than Newton, solely from a study of the motion of matter, was able to discover the Cause by which gravitational attraction is perform'd. Then, at the turn of the present century, a new state of matter was discovered, by physicists, on the Continent and in England, in evacuated tubes containing electrodes across which was applied a high voltage. These experiments produced visible, electrified rays. Astonishingly, the properties of these 'cathode rays'—their deflexion by electric and magnetic fields, for example—were *independent* of the chemical nature of the electrodes and the residual gas remaining in the tubes. Hence, concluded the Continental physicists, the cathode rays had nothing to do, directly, with either the cathode or the residual gas being connected in some way with the ether. Hence, concluded J. J. Thomson (b. 1856), the rays contain a *universal constituent of matter*, which he called the 'corpuscle'. Today we call it the 'electron'.

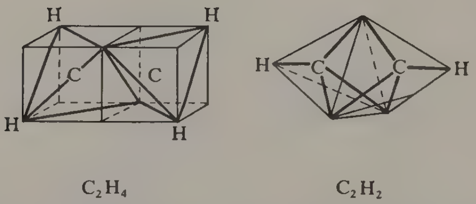


FIG. 5.4. Three-dimensional models of single, double, and triple bonds formed through the sharing by two tetrahedra of a corner, an edge, and a face, respectively, after van't Hoff (1874), for, respectively, $H_3C—CH_3$ (ethane), $H_2C═CH_2$ (tetrahedra placed in cubes; bonds not shown), and $CH≡CH$. Components of the double and triple bonds are 'bent'; i.e. the carbon valencies of multiple bonds do not point along the carbon–carbon axis.

At first it was not clear how to use the new particle in chemistry. A popular toast to 'JJ' at the annual Christmas party at the Cavendish Laboratory during this period was, 'Long live the electron! May it never be of use to anyone.' Thomson steadfastly insisted, however, that the origin of chemical attraction must be sought in the 'electronic structure' of matter.

Several years after Thomson's invention of the electron, and some thirty years after Mendeleev's discovery of the Periodic Law and the recognition, later, following the discovery of the noble gases, of the importance in chemistry of the number 8, the physical chemist Lewis (b. 1875) considered models of atoms in which the electrons in the atoms were placed at the corners of a cube. Lewis was handicapped in his speculations, however, by the fact that he did not know how many electrons there were in atoms. His 1902 model of helium, for example, contained eight electrons.

By 1914, particularly as a result of the work with X-rays of the physicist Moseley (b. 1887; d. 1915), the question as to how many electrons there are in atoms had been answered: the number of electrons in a neutral atom of an element is equal to the position of the element in the periodic table. Shortly thereafter, in 1916, in his classic paper 'The atom and the molecule', Lewis stated his *even-number rule*. Although half the atoms of the periodic table have an odd number of electrons, the number of electrons in compounds, Lewis pointed out, 'is in almost all cases an even number. . . . The extraordinary generality of this rule is shown by the fact that among the tens of thousands of known compounds of the elements under consideration [the non-transition elements] only a few exceptions [such as NO, with 15, and NO_2, with 23 electrons] are known'.

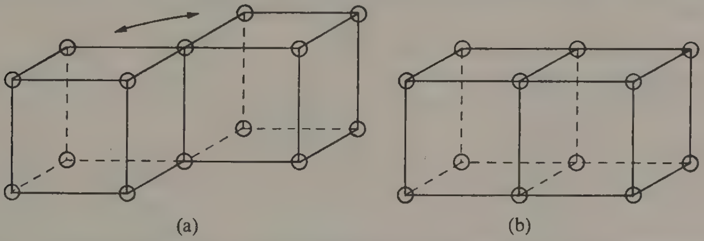


FIG. 5.5. Lewis's preliminary three-dimensional models of the electronic structures of single and double bonds, formed through the sharing by two cubes of an edge (a) and a face (b), respectively. The arrow indicates a pair of electrons that, from the modern point of view, would interfere with each other 'sterically'.

Initially, Lewis pictured a single bond as two cubes sharing an edge and a double bond as two cubes sharing a face (Fig. 5.5). There were serious difficulties with these cubical models, however. Neither model explained how hydrogen atoms might be bound to a carbon skeleton; the former model did not account for the experimentally observed nearly free rotation about single bonds (and, today, would be viewed as containing prohibitively large non-bonded interactions), and it was not clear how two cubes could be used to represent a triple bond.

'[When] we consider only known chemical phenomena, and their best interpretation in terms of atomic structure, we are led to assume a somewhat different arrangement of the group of eight electrons, at least in the case of the more nonpolar substances,' Lewis concluded. This arrangement is 'a group of eight electrons in which . . . *pairs* [of electrons] are symmetrically placed about the centre . . . [giving] identically the model of the tetrahedral carbon atom which has been of such signal utility throughout the whole of organic chemistry' (italic added).

Later, it was realized that, in relatively non-polar compounds, nitrogen and oxygen atoms, also, have their outer group of eight 'valence-shell' electrons arranged, like carbon, in pairs tetrahedrally placed about the centre of the atom. Unlike carbon, however, not all four of the tetrahedrally disposed valence-shell electron-pairs about chemically combined atoms of nitrogen and oxygen are necessarily shared. Some pairs, generally one pair in the case of a nitrogen atom and two pairs in the case of an oxygen atom, may be unshared. The facts of modern structural chemistry suggest that these 'lone pairs' occupy about as much space in an atom's valence-shell as do its bonding pairs. Thus, the ammonia molecule (NH_3) is not planar but pyramidal (H–N–H bond angles 107°; in the tetrahedral molecule CH_4 the H–C–H bond angles are 109·5°). Similarly, the water molecule (H_2O) is not linear but bent (H–O–H bond angle 104·5°), (Fig. 5.6).

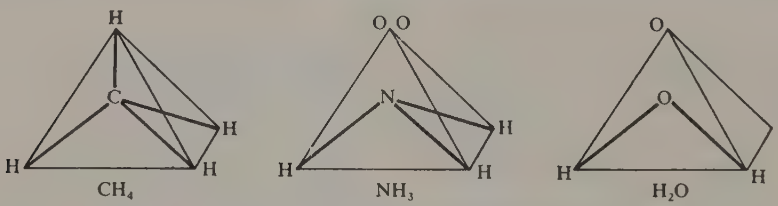


FIG. 5.6. The importance of lone pairs. An extension of the concept of the tetrahedral atom from carbon to nitrogen and oxygen. The molecules CH_4, NH_3, and H_2O are said to be 'isoelectronic' (*J. chem. Educ.* **43**, 170 (1966)).

Lewis's new conception—that the 'valence-strokes', as the solid lines in Crum Brown's graphic formulas were called, represent electron pairs—raised more questions that it answered. Why should electrons occupy space in pairs? And, if they should do so, why do not triplets and higher clusters occur? In short, what is the source of the special significance in chemistry of the number 2?

And there was another problem. A decade earlier Thomson had foreseen that, to have a static model of a stable atom, one must disperse at least part of the atom's charge. Thomson dispersed the atom's positive charge and imbedded in it his negative corpuscles, like plums in pudding—his famous, and generally accepted, (negative) plum- (positive) pudding model of the atom. To everyone's surprise,† however, Rutherford (b. 1871), in 1911, was led to suppose that, owing to the backward

† See Chapter 3, p. 54.

scattering of alpha particles by metal foils, most of the mass and all of the positive charge of an atom must be concentrated in a relatively small region of space, the 'nucleus'. Thus, in 1916 there was the question: What keeps Lewis's electrons out of Rutherford's nucleus ? The problem of explaining chemical attraction had become the problem of accounting for physical repulsion.

No satisfactory answer could be given to the previous question, nor to the question of electron pairing, in 1916. Nor is there today general agreement on how to best answer these questions—or even on whether these questions are the best ones to ask. Perhaps, to paraphrase Newton, chemists must learn from the phenomena of Nature how electrons behave in compounds, and what are the laws of their behaviour, before they ask what is the cause by which this behaviour is produced.

One law of electron behaviour in compounds might have been stated in 1916. This law is based on the observation that the valence-strokes of classical structural theory obey an exclusion principle: they never cross. In no instance has it been found useful to represent molecules by graphic formulae in which two (or more) valence-strokes cross each other. Taken with Lewis's identification of the valence-stroke, this non-crossing rule leads to the statement of a *chemist's exclusion principle: no more than two electrons may share the same region of space.*

Since all macroscopic samples of matter are composed of atoms, and since all atoms contain electrons, the chemists exclusion principle leads to a fundamental—that is to say widely applicable, that is to say very familiar, that is to say 'trivial'—*macroscopic exclusion principle*, which every child discovers at an early age: namely, that *two objects cannot be at the same place at the same time.*

The exclusion effect† was discovered by physicists nearly ten years later, by Pauli (b. 1900), after whom it is named. This *physicist's exclusion principle* (the Pauli principle) was generalized in 1926 by Dirac (b. 1902). It asserts that *electrons that have the same 'spin' cannot be at the same place at the same time.*

Since the 'spin' of an electron can have only *two* values, the chemist's exclusion principle is consistent with, if not immediately derivable from, the physicist's exclusion principle. The three exclusion principles—in order of increasing practical utility and decreasing mathematical precision, the physicist's exclusion principle, the chemist's exclusion principle, and everybody's macroscopic exclusion principle—appear to be mutually compatible, if not equivalent, statements.

† See Chapter 3, p. 67.

But exclusion principles alone do not account for matter's most salient feature: its bulk. Neither the chemist's nor the physicist's exclusion principle, by itself, yields a rationalization for Lewis's implicit assumption—which is consistent with everyday experience—that electrons (and therefore atoms) have, in effect, a *finite size*—an effective 'radius'—which prevents electrons (and atoms)—and, therefore, matter generally—from collapsing to a point.

To account for the bulkiness of matter and, generally, to facilitate the application of the concept of spatial exclusion to complex chemical systems, not to mention its application to ordinary, macroscopic systems, another postulate is useful, perhaps necessary: the *postulate of finite size*. Indeed, it would seem pointless to apply a principle of spatial exclusion to 'real' point particles.

To describe the spatial relationships of objects to each other, in particular, their distances of closest approach, it is convenient to assume that, for instance, a table has certain linear dimensions, whether it is in the living-room of a house or in the dining-room. Similarly, in structural theory it is useful (in a first appoximation) to associate a certain radius with, for example, a nitrogen molecule, whether it is in the gaseous, liquid, or solid phase; or with a sodium ion, whether it is in a crystal of sodium chloride or sodium bromide; or with an electron pair, whether it is between two carbon atoms in ethane, hexachloroethane, or diamond. Of course, since all substances are compressible, the effective size of an object varies with external circumstances. In particular, the effective sizes of electrons in chemical compounds are moderately sensitive functions of the charges on, and the sizes of, the atomic cores to which they are bound.

In physics, the postulate of finite size is expressed in the form of a differential equation discovered by Schrödinger (b. 1887) in 1926.† In effect, it is assumed that, after de Broglie (b. 1892), electrons have momenta that are inversely proportional to their effective sizes, or wavelengths. As an electron cloud contracts about a nucleus, in an effort to diminish the electrostatic energy of the system, the momentum and associated kinetic energy of the system increase. Eventually, a balance is reached between the increase in the kinetic energy of the system and the decrease in its potential energy. That, from the viewpoint of electron physics, is the reason why Lewis's electrons stay suspended—or extended—in space. They resist compression, owing to their wave-like character.

Molecular models that incorporate the insights into structural chemistry of Frankland, Kekulé, van't Hoff, and Lewis; the results on the

† See Chapter 3, p. 69.

structure of atoms of Thomson, Rutherford, and Moseley; and the views regarding the nature of electrons of de Broglie, Schrödinger, Pauli, and Dirac, are shown in Fig. 5.7, for ethane, ethylene, and acetylene. Each sphere represents one valence-shell electron-pair; i.e. each sphere corresponds to one of the valence-strokes in the graphic formula for the molecule.

Still, the question remains: Why in compounds should two electrons share the same region of space? As Lewis himself acknowledged, 'If we are to speak at all of a force between the two electrons of a pair, it is on the whole a repulsive force.' Yet electrons do cluster about nuclei, in free atoms, in direct opposition to electron–electron repulsion, owing to nucleus–electron attractions. Additional electron clustering occurs in molecules, owing to the particularly large nucleus–electron attractions that exist in those regions—the 'bonding regions'—between two or more nuclei. Spatial pairing of electrons in molecules occurs not because of a special force between electrons. It occurs because nucleus–electron attractions overwhelm electron–electron repulsions.

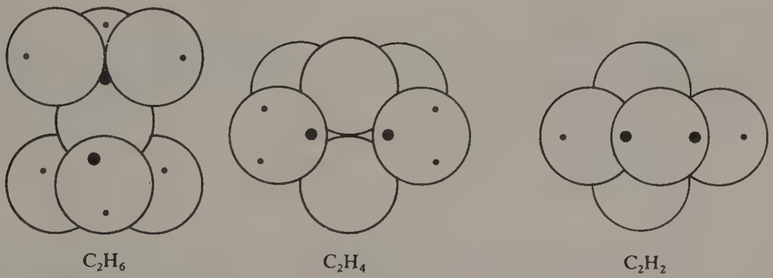


FIG. 5.7. Tangent-sphere representation of ethane, ethylene, and acetylene. Large circles represent valence-shell electron-pairs (the valence-strokes in Crum Brown's graphic formulae). Small solid circles represent protons (the cores of hydrogen atoms). Larger solid circles represent carbon-atom cores (carbon nuclei plus inner-shell electrons). Each carbon core is surrounded by four tangent spheres.

For some molecules, it is possible to achieve a normal, tetrahedral arrangement of electrons about each atomic core for each spin-type separately without having the electrons of one spin-type share fully the space occupied by electrons of the other spin-type. For the oxygen molecule, for example, Linnett (b. 1913) has pointed out recently that, owing to an improvement in repulsions between electrons of opposite spin, the electron arrangement shown schematically in Fig. 5.8 (b) is a better arrangement than the one shown in Fig. 5.8 (a) .The latter arrangement (b) is nothing more (or less) than Lewis's arrangement of electrons

(5.5b) refined to allow for the property of electron spin. Fig. 5.8 (b) represents the mutual disposition of the electrons in low-lying excited state of the oxygen molecule. In its ground state, the oxygen molecule has 7 valence-shell electrons of one spin-type and 5 of the other. In van't Hoff–Lewis–Linnett theory, they would be arranged as shown in Fig. 5.8 (c).

The structural theory developed by van't Hoff, Lewis, and Linnett focuses attention on compounds of carbon, nitrogen, oxygen, and fluorine. But, 'How do we know', asked Wigner (b. 1902), 'that, if we made a theory which focuses its attention on phenomena we disregard and disregards some of the phenomena now commanding our attention, that we could not build another theory which has little in common with the present one but which, nevertheless, explains just as many phenomena as the present theory?'

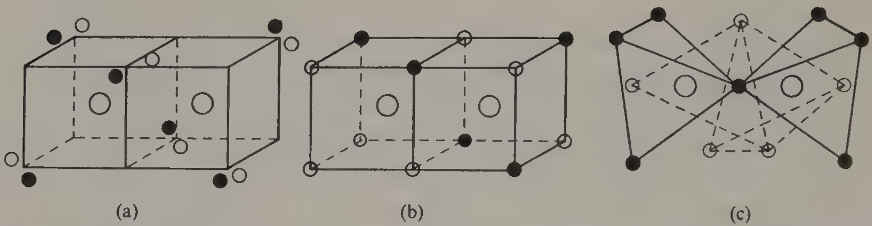
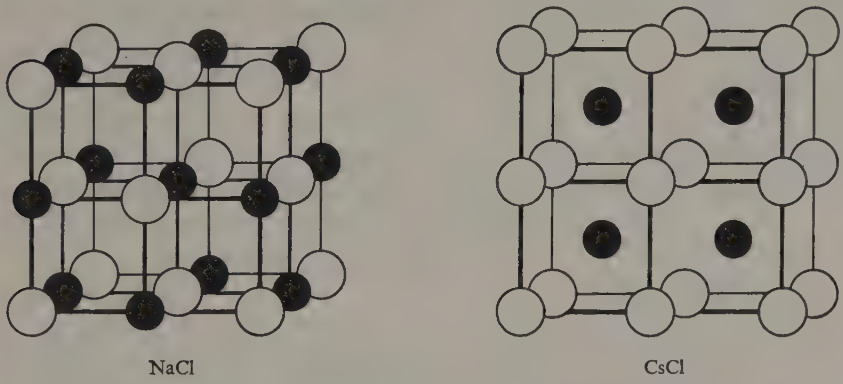


FIG. 5.8. Models of the electronic structure of O_2, the oxygen molecule. Small filled circles represent valence-shell electrons of one spin. Small open circles represent valence-shell electrons of the other spin. Larger open circles represent oxygen-atom cores (oxygen nuclei plus inner-shell electrons). Each oxygen core is surrounded by four, tetrahedrally arranged electrons of each spin type. In model (a) electrons share regions of space in pairs; the two 'spin-sets' are 'coincident'. In model (b) the spin-sets are somewhat anticoincident. From the viewpoint of electron–electron repulsions, model (b) is a better model than model (a). Evidently model (c) is still better, for in its ground state the O_2 molecule has 7 valence-shell electrons of one spin and 5 of the other. Like models (a) and (b), model (c) places four electrons in the bonding region, which accords with the physical properties of O_2.

Indeed, in 1912, through the use of X-rays, the Braggs, father (b. 1862) and son (b. 1890), discovered that, as Barlow (b. 1845) had anticipated in 1883 from studies of the shapes of crystals, sodium chloride has the structure shown in Fig. 5.9 (a). Each sodium atom is surrounded by 6 chlorine atoms and each chlorine atom is surrounded by 6 sodium atoms. On the other hand, in caesium chloride (Fig. 5.9 (b)) each caesium atom is surrounded by 8 chlorine atoms and each chlorine atom is surrounded by 8 caesium atoms. Similarly, in crystals of the difluorides of the Group II metals, BeF_2, MgF_2, and CaF_2, the 'coordination numbers' of the atoms increase—from 4 through 6 to 8 for the metal atoms, from 2 through 3 to 4 for the fluorine atoms—as one passes from the low-atomic-weight element beryllium through magnesium to the higher-atomic weight element calcium. All these compounds dissolve in water and give electrically conducting solutions, owing to the presence of charged atoms or 'ions', which Arrhenius had suggested in 1884. The differences in their crystal structures might be due, suggested the younger Bragg, Goldschmidt (b. 1888), and others, to differences in the *sizes* of the ions. The larger a positive ion, the larger the number of negative ions that can be packed about it without creating a hole so large that the positive ion 'rattles'—which would be 'bad' electrostatically. The 'ion-packing' model of sodium chloride is shown in Fig. 5.10. In such models, there are no molecules. *The whole crystal is one large molecule.*


NaCl

CsCl

FIG. 5.9. The structures of sodium chloride and caesium chloride.

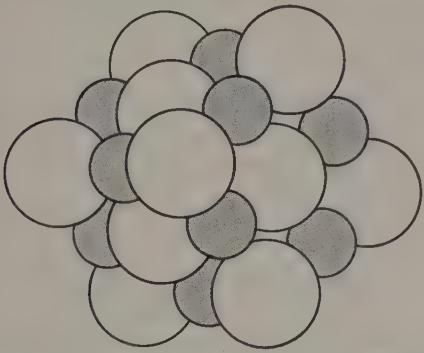


FIG. 5.10. Tangent-sphere model of sodium chloride, after Barlow. Small spheres represent Na^+ ions, large spheres Cl^- ions.

At first glance, the ion-packing model of crystals looks very different from the electron-sharing model of molecules. True, there seem to be large differences in terminology. In the ionic model of crystals, one speaks of 'non-directional bonding' and of 'shared corners', 'edges', and 'faces' of 'polyhedra of negative ions' packed about smaller positive ions. In the shared-electron model of molecules, one speaks of 'directional bonding' and of 'single', 'double', and 'triple' bonds between 'atoms'.

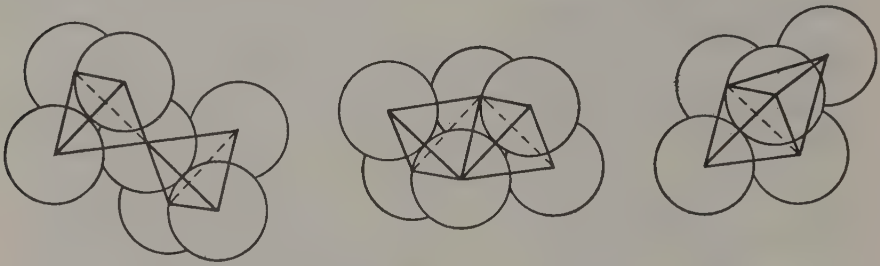


FIG. 5.11. The sharing of a corner, an edge, and a face by a pair of tetrahedra, from Pauling's work on the structure of complex ionic crystals. Each sphere in this figure represents a negative ion (e.g. O^{2-}, F^-). Smaller positive ions (not shown) occupy the tetrahedral interstices between the larger negative ions; thus, each positive ion is surrounded by four tetrahedrally arranged negative ions, some of which it shares with an adjacent positive ion. The tetrahedral arrangement is produced solely by packing considerations: that is to say, by electrostatic forces and the postulate of finite size.

These differences are more apparent than real, however, once the wave properties of electrons *and* the exclusion principle are fully represented in the electron-sharing model. This step was first taken about ten years ago. The results have been shown in Fig. 5.7. Fig. 5.11 is taken from a 40-year-old classic by Pauling (b. 1901) on the structures of complex ionic crystals. It shows two corner-, edge-, or face-sharing tetrahedra of negative ions; each tetrahedron of negative ions surrounds a smaller, positive ion, not shown. The analogy with the 'tangent-sphere' representation of molecules, Fig. 5.7, is apparent. The negative 'ions' in the tangent-sphere representation of molecules are the electron-pairs of the carbon–carbon bonds and the protonated electron-pairs of the carbon–hydrogen bonds; the positive 'ions', which are much smaller than the negative 'ions' and, as in Fig. 5.11, are surrounded by four negative ions, are the atomic cores (carbon nuclei plus inner-shell electrons). With this understanding, most of the theory of 'ionic' crystals can be applied, without change, to ordinary 'covalent' molecules. For molecules, as for crystals, we have an ion-packing model. *A molecule is merely a very small crystal.*

These ion-packing models may be viewed as (possibly) crude but practical representations of the implications for chemistry of the wave-like properties of electrons and the exclusion principle. In effect, chemists have been articulating, slowly but surely, for over a century, without realizing it, a *mechanics of the exclusion principle*. Unlike ordinary mechanics, this chemical mechanics has the peculiar, but to chemists familiar, feature that it produces *stable structures*. The theory of these structures—structural theory—is a chemical invention. The supporting evidence, typically, while voluminous, is indirect. As yet, there is little independent evidence or theoretical support—that is to say, little evidence from physics—for some of the most fundamental features of the simplest graphic formulae. Chemists are not quite sure, therefore, if they should believe them. But they cannot do without them.

Useful as it is, there is a great gap in structural theory. Most of the elements of the periodic tables are metals, yet chemists do not have a satisfactory theory of the metallic bond. Perhaps it is merely a coincidence that the arrangements of the atoms in many metals are the same as the arrangements of the electropositive atoms in many salts. The arrangement of the calcium atoms in one form of calcium metal, for example, is the same as the arrangement of the sodium ions in sodium chloride (Fig. 5.9 (a)), which is the arrangement, also, of the calcium atoms in calcium fluoride (CaF_2). Such might be the case, however, if, in metals, electrons played the role of negative ions. This simple model accounts for the structures of many metals and for some of their more familiar chemical properties. But whether it is a generally useful model of metals remains to be seen.

Selected References

Benfey, O. T. (editor). *Classics in the theory of chemical combination.* Dover, New York (1963).

Bent, H. A. Tangent-sphere models of molecules. *J. chem. Educ.* **40**, 446, 523 (1963); ibid **42**, 302, 348 (1965); ibid **44**, 512 (1967); The tetrahedral atom. *Chemistry* **39**, 8 (1966); ibid **40**, 8 (1967).

Goldschmidt, V. M. Crystal structure and chemical constitution. *Trans. Faraday Soc.* **25**, 253 (1929).

Greenaway, F. *John Dalton and the atom.* Cornell University Press, Ithaca, New York (1966).

Holmyard, E. J. *Makers of chemistry.* Oxford University Press (1962).

Ihde, A. J. *The development of modern chemistry.* Harper and Row, New York (1964).

Lewis, G. N. *Valence and the structure of atoms and molecules.* Dover, New York (1966); The atom and the molecule. *J. Am. chem. Soc.* **38**, 762 (1916).

LINNETT, J. W. *The electronic structure of molecules. A new approach.* Methuen, London (1964); A modification of the Lewis–Langmuir octet rule. *J. Am. chem. Soc.* **83**, 2643 (1961).

MARGENAU, H. The exclusion principle and its philosophical importance. *Phil. Sc.* **11**, 187 (1944).

PALMER, W. G. *A history of the concept of valency to* 1930. Cambridge University Press (1965).

PAULING, L. *The nature of the chemical bond*, 3rd edn. Cornell University Press, Ithaca, New York (1960); The principles determining the structure of complex ionic crystals. *J. Am. chem. Soc.* **51**, 1010 (1929).

THOMSON, G. P. *The inspiration of science*. Oxford University Press (1961).

WELLS, A. F. *Structural inorganic chemistry*. Oxford University Press (1962).

6 *Biochemistry*

'The more we know of any branch of science, the less is the compass into which we are able to bring its principles, provided the facts from which they are inferred be numerous. Because, in an advanced state of knowledge, we are able to reduce more of the *particular* into *general* observations; whereas, in the infancy of a science, every observation is an independent fact; and, in delivering the principles of it, they must all be distinctly mentioned; so that though a *selection* may be made, a proper *abridgement* is impossible.' Joseph Priestley, 1772.

BIOCHEMISTRY is defined (by the *Oxford English Dictionary*) as 'the chemistry of living organisms; biological, physiological or vital chemistry'. This definition, however, does not altogether catch the spirit of the subject. I should rather say that the biochemist seeks to explain biological phenomena in chemical terms. Explanations in biochemistry are couched in terms of the reactions (and interactions) between molecules. A typical problem is 'how does the food that we eat give us the energy that we need?', and the explanation given in textbooks of biochemistry consists essentially of a description of the various chemical reactions that take place. Prediction is pretty limited, especially in multicellular organisms with their complex reciprocal interactions.

To return to the definition offered above, the peculiar nature of biochemistry lies in the attempt to answer biological questions by chemical methods: the problems are those of one discipline, the tools are those of another. The biochemist is profoundly (sometimes unconsciously) affected by the unifying concepts and guiding principles of biology, such as evolution and natural selection, the wholeness and purposiveness of organisms, or the stability of internal environment, and yet he must master the chemistry that he uses. It is now taken for granted that there are no 'new' forces in biochemistry, that cells (like molecules) are ultimately explicable in terms of the electrostatic forces between protons and electrons.†

It will be convenient to start a discussion of some leading ideas in biochemistry by outlining some of our ideas on enzymes, for the study

† See Chapter 5, p. 136.

of enzymes plays an exceedingly important part in biochemistry. It can easily be seen why this is so. The biological events that fall within the province of biochemistry are mostly chemical reactions. Now it turns out that practically all the chemical reactions in living cells are catalysed by enzymes. Enzymic catalysis is thus the basic chemical manoeuvre of life. Hence catalysis is the main theme of this chapter. We shall start with some features of enzymes as catalysts, and then go on to the nature of enzymes and thence to metabolism, which is essentially organized catalysis. I shall try to illustrate the point that biochemistry has been advanced as much by the development of techniques as by the elaboration of concepts.

The reactions catalysed vary much in their nature; some proceed quite rapidly even without added catalysts, but a few are quite unexpected from the chemist's viewpoint. Two reactions of a 'simple' (anyhow, small) molecule illustrate well this divergence. Carbon dioxide reacts rapidly with water even without an added catalyst, yet there is an enzyme that catalyses this reaction in red blood cells. In the other reaction (the chemically unexpected one) carbon dioxide reacts with ribulose diphosphate to give phosphoglyceric acid (p. 149). On this last reaction depends all plant (and hence animal) life, for it is the first stage in the reduction of atmospheric carbon dioxide to carbohydrate, and it proceeds on the scale of 6000 tons per second.

The only large class of non-enzymic reactions in cells is ionization reactions. In these reactions a proton is transferred from one molecule to another, and this transfer is often very rapid.

The concept that reactions in cells are enzymic reactions could naturally not be established by one (or a few) experiments, but has grown gradually as the result of innumerable studies. This concept is now well-grounded; there are roughly one thousand examples known, and probably not more than another few thousand yet to be discovered. It is the general validity of this concept that makes the study of enzymes such an important part of biochemistry.

Enzymology is a twentieth-century branch of science. The stage was set at the close of the last century with the demonstration by Edouard Büchner that the juice obtained from yeast cells by grinding them and squeezing with a press could ferment sugar. This may sound a somewhat matter-of-fact observation, but it was both revolutionary in its time and had far-reaching consequences. It was revolutionary because the study of fermentation had been a major topic for many years. Controversy had raged around the question of whether yeast was alive, and about the relationship between yeast and fermentation. Liebig was reluctant to admit that yeast consisted of living cells. Pasteur eventually marshalled overwhelming evidence that yeast cells could multiply. Now the yeast cells were recognized by their action in bringing about alcoholic fermentation, i.e. the conversion of glucose into ethanol and carbon dioxide. Pasteur, however, identified the agent with the action: he defined alcoholic fermentation as ‘the fermentation that sugar undergoes under the influence of a ferment that bears the name of brewer’s yeast’. Yet Büchner later demonstrated that the juice that he had obtained was free from yeast cells and still was very active in bringing about fermentation. Hence fermentation could, after all, take place ‘without life’—but of course the active agents (enzymes) had come from living cells. Büchner’s discovery, made seventy years ago, paved the way for study of the chemical agents that bring about a vital act. The historical importance of fermentation is acknowledge in the term enzyme (from *ἐύζυμος*, leavened); the English word *enzyme* was taken from the German, but paradoxically ‘ferment’ is still sometimes used in German.

Some extracellular enzymes (unorganized ferments) had been known previously; it was the existence of intracellular enzymes that was novel at the beginning of the century. Their importance was soon realized by some workers. Thus Hofmeister (in 1901) predicted that sooner or later each vital chemical reaction would be found to be brought about by a characteristic enzyme. Enzymes, Hofmeister said, were the essential chemical tools of the cell, and he emphasized that an ordered sequence of reactions signified chemical organization in the cell.

What are the peculiar advantages of enzymic catalysis that are utilized in the cell? The main ones can be classed as ‘specificity’ and ‘speed’, and they are related. Quantitative reactions of organic compounds are none too common. Side-reactions usually intervene. Competing reactions occur at comparable rates. In an enzymic reaction, one mode of reaction is hastened but the others are not. Side-reactions are thus avoided: instead of several ways over the hill there is one tunnel through it. Enzymes are usually specialized catalysts. There are obvious social, and physiological, parallels. An organized state demands workers who are specialists. A multi-cellular organism has cells that are specialized. Similarly, the specialized function of the enzyme fumarase, for example, is to convert fumaric acid into malic acid.

The other feature of enzymic catalysis is speed. Enzymic reactions may proceed a thousand million times as rapidly as their non-enzymic counterparts. This difference corresponds, on a time-scale, to the difference between seconds and centuries. Many reactions in living cells take place in a few seconds, and they do so in neutral solution. Thus the lifetimes of fumaric and malic acids in the kidney is only a few seconds. The non-enzymic reaction in neutral solution has to be carried out at a temperature of about 200°C for the rate to be conveniently measureable. The reaction can be carried out at a lower temperature, but then strongly acidic solutions must be used.

It is well known that cells are organized in space, i.e. morphologically. But the reactions in cells are organized in time. This is expressed metaphorically in the term 'metabolic pathway'. The pathway is a spatial concept but the term refers to temporal organization. The metabolic pathway is a chain of reactions. The reactions take place successively; for example, glucose is converted into glucose-6-phosphate, and then this into fructose-6-phosphate, and so on. There are a dozen successive reactions in this pathway (glycolysis) and each reaction is catalysed by a separate enzyme. If one enzyme were missing, the corresponding reaction would not proceed at anywhere near an adequate rate, and so the whole pathway would be inoperative. It is essential that the substrates should undergo specific reactions, for example that only one out of the five hydroxyl groups in glucose should be phosphorylated. This specificity is achieved by enhancing the reactivity of one particular hydroxyl group rather than by decreasing the reactivities of all the others. Without the benefit of enzymes, the chemist has (in general) to perform the relatively clumsy manoeuvre of protecting the other reactive groups in order to phosphorylate only one of them.

Some features of enzyme reaction have now been mentioned, but nothing has so far been said about the nature of enzymes. In fact, knowledge of the chemical nature of enzymes lagged far behind knowledge of their catalytic properties. Much was known about enzymes as catalysts in the 1930s, but little about enzymes as molecules. Much effort was devoted to the purification of enzymes and considerable progress was made, and yet their nature was not discovered. This was partly because the enzymes being studied were active in such dilute solution that the sensitivity of chemical tests (for protein, carbohydrate, etc.) was inadequate. The crystallization of urease by Sumner in 1926 is now seen as the first step in the right direction. This was followed by the isolation of several other crystalline enzymes by Northrop and Kunitz, and by the intensive study of their properties. It turned out that the enzymes were proteins. This was a great step forward. However, these were mostly extracellular enzymes. Although they are far from being typical enzymes, extracellular enzymes were recognized early, and indeed it is still true that more is known about them. Dixon and Webb comment as follows: 'In fact the serious purification of intracellular enzymes did not begin until 1937, despite the fact that comparatively few enzymes occur naturally outside living cells.' In the last thirty years, however, several hundred intracellular enzymes have been isolated. These, too, are proteins. In fact, Dixon and Webb suggest that the most satisfactory definition of an enzyme is 'a protein with catalytic properties due to its power of specific activation'. It is the chemically diverse nature of the side chains (see p. 146-7) that enables proteins to function as enzymes. The conclusion that enzymes are proteins, together with the conclusion that chemical reactions in cells are catalysed by enzymes, comprise two of the most important concepts in biochemistry.

Some enzymes are not composed exclusively of protein. Enzymes may contain a non-protein prosthetic group that plays an important part in catalysis. Thus, the aptly-named 'old yellow enzyme' contains riboflavin-5′-phosphate as prosthetic group. In fact, the presence of a non-protein component is helpful in proving that the protein component is essential. Theorell found that the prosthetic group could be separated by dialysis in acid solution; the separation abolishes the catalytic activity. But activity can be restored by addition of a stoicheiometric amount of the riboflavin derivative; there are simultaneous changes in colour and fluorescence. Riboflavin is a vitamin (see p. 148). Prosthetic groups cannot always be clearly differentiated from coenzymes; both terms refer to the non-protein conserved components of catalytic systems. A prosthetic group is part of a conjugated protein, i.e. a protein not made up wholly of amino acids but also containing another kind of structure, this latter being the phosthetic group. Coenzymes will be discussed in more detail later; we must now return to the implications of the conclusion that enzymes are proteins. To understand these implications, it is necessary to mention a few of the concepts and techniques that have been most important in the development of protein chemistry.

The hypothesis that proteins contain chains of amino acid residues, linked by the peptide bond, was put forward by Fischer and by Hofmeister (independently) at the beginning of the twentieth century; this is a good starting-point, both logically and chronologically. One line of evidence for this hypothesis was that proteins, although built up from amino acids, contained relatively few acidic and basic groups. This fact is explained by the peptide bond, which is formed by the mutual 'annihilation' of an acidic and a basic group. Another line of evidence is that proteolytic enzymes attack synthetic peptides; compounds known to contain peptide bonds are hydrolysed by enzymes that break down proteins. Reliable methods for determining the size of protein molecules were not available early in this century, but it seemed clear that their size was large.

Some of the concepts that have been important in the development of protein chemistry were derived from macromolecular chemistry. This term, in fact, embodies an idea that was novel and not readily accepted. When Staudinger suggested, in the 1920s, that molecules could be so large that they contained 10^5 or 10^6 atoms, this was quite an unexpected proposal. The more complicated molecules being studied then contained around 100 atoms. Larger particles were regarded as belonging to the colloidal state, and as being aggregates built up from a large and indefinite number of small molecules. Staudinger's ideas gained eventual acceptance, and formed the basis of polymer science. It was only with the development of polymer chemistry that it was found, for example, that groups at the ends of long chains had normal reactivity. In fact, we now regard macromolecules as being little different from small molecules, apart from their size.

At about the same time that the foundations of macromolecular chemistry were being laid, Svedberg developed the ultracentrifuge and showed that several proteins were monodisperse, i.e. that the solutions contained particles of uniform size. The proteins behaved, in fact, as if they were large molecules, containing about 10^4 atoms. This work confirmed the conclusions reached by Sorensen and by Adair, based on measurements of osmotic pressure. Much physico-chemical work on proteins in solution, and X-ray work on protein fibres, confirmed the general idea of proteins as macromolecules (see Chapter 7).

Detailed knowledge of the structure of proteins has come only relatively recently; this knowledge could not have been gained without the development of improved analytical techniques. Cells contain so many compounds that the biochemist invariably starts work with a formidable mixture. The separation of one compound from all the others present is thus likely to be difficult. Moreover, the protein chemist has several analytical problems in succession. First, the protein to be studied has to be separated from the several hundred other proteins generally present. Then the amino acids have to be separated to determine how many members of each kind of amino acid are present. Last (but not least—in fact most) the peptides produced by partial splitting have to be separated. The discovery of partition chromatography by Martin and Synge has had an enormous effect on biochemistry in general and protein chemistry in particular; it was this technique that Sanger exploited so successfully in his elucidation of the structure of insulin. Most experimental work in biochemistry now makes use of one or more of the following analytical techniques: paper (or thin-layer) chromatography, ion-exchange chromatography, gas chromatography, and electrophoresis. Chromatography is a method of separation introduced by the Russian botanist Tswett early in the century. In this method a mobile phase moves over a stationary phase; the rate at which a substance moves depends on its relative affinity for these two phases, and substances are separated when they move at different rates. For instance, in paper chromatography, a solvent (the mobile phase) flows down a sheet of filter paper (the stationary phase), and the separated substances are then detected by a colour reaction. It is interesting that several of the methods are quite simple, and could even have been discovered in the last century. It is a problem for the historian of science to uncover the reasons why several powerful analytical techniques should have become available at much the same time.

The most important technique in the final stages of studying a protein is X-ray diffraction, because this is the only way of determining the positions of the atoms in space, i.e. the three-dimensional conformation of the molecule (Chapter 5). The biological role that a protein plays is governed by its conformation, and this in turn is governed by the amino acid sequence. In most soluble proteins the peptide chain (or chains) is folded several times, so that the shape of the molecule is (very roughly) spherical. The irregular surface is studded with knobs and dents and these presumably give rise to 'sticky patches'. The stickiness is quite selective, and this is important in problems of 'recognition'. Most proteins contain several chains; each chain is coiled, and the coiled chains interact. After having been manufactured in the cell, the coiled chains belonging to one protein must recognize each other and 'bed down' together.

We now come back to enzyme action. In the earlier work on enzymes, an important landmark was the idea that the enzyme forms a complex with the substrate. This is now taken for granted, but it was by no means obvious at the beginning of the century. The hypothesis was put forward to account for the variation in rate with concentration of substrate. Establishing the quantitative laws governing catalysis by enzymes is now an important branch of biochemistry. Much of the groundwork was carried out before highly purified preparations of enzymes were available. Studies with inhibitors, for example, can often be carried out directly on tissue extracts. Indeed, this may be the most useful approach in metabolic studies, and it has often been fruitful in the past. The quantitative study of multi-enzyme systems is difficult; only simplified model situations have been analysed algebraically, and more detailed work relies on numerical analysis.

The concept of an enzyme–substrate complex leads us to the consideration of 'active sites'. The active site is generally taken to mean that part of the enzyme which is directly concerned in its catalytic action. A small molecule such as glycerol can interact with only a fraction of the whole molecule of a protein. The atoms in the enzyme which are in contact with those of the substrate are clearly important; many other parts of the enzyme no doubt play a supporting role. Much knowledge of active sites has been gained by chemical methods. For example, the glycolytic enzyme aldolase interacts with its substrate by forming an imine bond. Horecker has shown that the imine can be converted from a reactive intermediate into a stable (inactive) compound by reduction. The imine is formed from the side-chain amino group of lysine. Determination of the amino acid sequence around this lysine residue characterizes one part of the active site. Moreover, it is easy to understand the chemical basis for this feature of the action of aldolase. The formation of an imine activates the substrate; imines (in the protonated form) are more reactive than ketones. Other groups in the enzyme certainly take part in the catalysis, but since the conformation of aldolase is not known, we are ignorant about how the groups cooperate.

The first detailed hypothesis about the cooperation of groups in an active site comes from the X-ray studies of Phillips and his colleagues on lysozyme. One of the most interesting findings is the large number of interactions between enzyme and substrate. An important feature of the hypothesis is the distortion of the substrate as it is bound to the enzyme. This is one of the oldest ideas about how enzymes act, and now (in this instance) it is receiving experimental support. This quite new knowledge promises to place our understanding of enzyme action on a much firmer basis than before. It is probably true to say that we may now hope to find out how enzymes *do* act, whereas before we were trying to see only how enzymes *could* act.

We may summarize in a general way some of the current ideas about enzymic catalysis as follows. The enzyme–substrate complex contains a number of weak bonds. Several weak bonds can be rapidly made and broken, as is necessary for catalysis, and they permit accurate placing of the substrate. The enzyme preferentially stabilizes the transition state

(Chapter 7) of the reaction, i.e. the enzyme binds the transition state more firmly than it binds the substrate or the product. The enzyme distorts the substrate, since bond angles and distances in the transition state differ from those in the ground state. The conformation of the enzyme will be altered to a greater or lesser extent by the interaction.

It is becoming somewhat clearer now why enzymes are large molecules. Only large molecules are capable of multiple interactions, and moreover it is doubtful whether a small molecule could hit the right balance between rigidity and flexibility.

The importance of weak bonds in enzymes is but one example of their wide importance in the structures of proteins, nucleic acids, and other ordered macromolecules. Moreover, as we grope our way upwards from macromolecules to membranes and organelles we shall have to pay more and more attention to weak bonds. Weak bonds can be classified as polar or non-polar; the dualistic theory of Berzelius (1811) reappears in a new context. Hydrogen bonds are polar weak bonds, and arise because the hydrogen atom has only one electron. Thus hydrogen bonding between two molecules of water is due to the attractive force between the proton covalently bound to the oxygen atom of the first molecule and the unshared electrons of the oxygen atom of the second molecule. The strong hydrogen bonding in water is an important factor in governing the properties of compounds in aqueous solution; indeed, it can be regarded as responsible for non-polar interactions in aqueous solution. Non-polar interactions (hydrophobic bonds) are largely responsible for the compact shape of globular proteins. The non-polar side chains seek an escape from their watery domain and come together in the interior of the molecule. Hence the chains fold, and if there are still many non-polar residues that cannot be tucked inside, two (or more) folded chains may aggregate.

The next stage of organization above that of individual enzymes is represented by multi-enzyme complexes. An example is pyruvate dehydrogenase. The multi-enzyme complex from bacteria has a particle weight of about five million. This complex contains 25 molecules of enzymes (16 molecules of pyruvate decarboxylase, 8 molecules of dihydrolipoic dehydrogenase, and 1 molecule of lipoic reductase—transacetylase), and each of the three different kinds of enzyme in the complex contains its own coenzyme. The complex is thus quite highly organized.

The enzyme just mentioned, pyruvate decarboxylase, requires thiamine pyrophosphate as coenzyme. Thiamine is a vitamin; one of the achievements of twentieth-century biochemistry is the discovery of vitamins and (in some cases) the ability to understand their role in chemical terms. The development of the concept of deficiency diseases (e.g. beri-beri) is too well known to need recapitulation, but it will be recalled that one of the main leads came from the experiments of Hopkins (in 1912) on the growth of rats that were fed controlled diets. Many vitamins were discovered during the next thirty years; they were found to be functionally (and chemically) diverse. The role of thiamine is relatively well understood. The pyrophosphate is the coenzyme for several enzymic reactions. These reactions share a common feature: they all consist of fissions of a bond joining a carbon atom to a carbonyl group. It is quite clear what, in chemical terms, the thiamine pyrophosphate is doing, and how it does it. A disease (beri-beri) can thus be related in a general way to the role of a vitamin derivative as a coenzyme. However, the detailed relationship between the metabolic lesion and the development of certain symptoms is far from clear.

We have now said something about enzymes and must approach metabolism, a complex interlocking network of enzymic reactions. The detailed knowledge of metabolism that we now possess represents perhaps the most solid achievement of biochemistry and forms the main part of the contents of textbooks of biochemistry. Biochemists study metabolic reactions at a variety of levels, and by a variety of techniques. The level may be that of the whole organism, or that of a slice or mush consisting largely of intact cells, or that of the cell-free extract. The techniques of spectroscopy, of manometry, and of the use of isotopes as tracers are widely used, and here too progress has been greatly influenced by the development of techniques. The use of isotopes as tracers has been particularly fruitful; and the fact that tracers can be easily used in studying the metabolism of living cells is a great advantage. The development of this technique is largely due to de Hevesy. In an early biological application (1923), he found that roots, after taking up a radioactive isotope of lead, lost the radioactivity on exposure to solutions containing ordinary (unlabelled) lead ions; this result established that ions migrated out from plant roots. One general idea that has emerged from studies with tracers is expressed in the title of Schoenheimer's book: the dynamic state of body constituents. This work was mainly concerned with the proteins in animal tissues, and led to the concept that macromolecules might be continually being broken down and built up again within cells. Thus stability is maintained by flux. This concept has been most influential, but in fact much remains to be learned about the extent and control of protein turnover.

One of the most spectacular applications of tracers was the work of Calvin on the path of carbon in photosynthesis. This work was started in 1945 when the radioactive isotope (^{14}C) of carbon became available and has led to the formulation of a photosynthetic carbon reduction cycle consisting of about a dozen enzymic reactions. The key compound is phosphoglyceric acid; its identification depended on the use of tracers combined with the use of paper chromatography. This combination of methods has subsequently proved very powerful in other investigations of metabolic pathways. The procedure in Calvin's experiments was to expose unicellular green algae to isotopically labelled carbon dioxide for a few seconds and then to plunge them into hot methanol to arrest the reactions. The radioactive compounds formed from the carbon dioxide were identified by paper chromatography followed by radioautography. The fate of the labelled atom can be followed, when the organism is growing steadily, by interrupting the experiments at various intervals of time. At the shortest times, the products first formed are the most highly radioactive, and hence the pathway is elucidated.

The nutritive requirements of organisms constitute a valuable method for investigating their metabolism, and led (as mentioned above) to the discovery of vitamins. This approach is widely used in work with microorganisms. Thus the studies of Beadle and Tatum on mutants of the mould *Neurospora* led to the 'one gene, one enzyme' hypothesis; although these ideas are described in another context (Chapter 7), in fact they now permeate biochemical thought.

We come now to one of the most striking findings of biochemistry: the unity underlying the diversity in biology. Variety is perhaps the most obvious characteristic of living things. Certain unifying concepts were already established in the last century, notably the cell theory. Similarities in composition, such as the presence of proteins in all cells, were also known. But what we are now concerned with is uniformity in processes. The first example of a common metabolic sequence was glycolysis. Harden, in 1926, pointed out the similarity in the carbohydrate metabolism of yeast and muscle, and this was reaffirmed by Meyerhof in 1937 when more of the reactions had been established. Alcoholic fermentation in yeast and sugar breakdown in muscle seem diverse enough happenings, but chemically they have much in common; in fact, nearly all the steps are the same. Here is an example of unity underlying diversity. This example concerns energy production, and many of the main features of mechanisms that produce energy turn out to be similar in all organisms. The key compound in all cells is adenosine triphosphate (ATP). The anaerobic mechanism for energy production in animal, plant, and most bacterial cells is glycolysis: glucose is converted into lactic acid, and ATP is formed. One of the difficulties in elucidating the pathway of glycolysis was that the 'objective', namely ATP production, was far from obvious. The complete conversion of glucose into carbon dioxide and water requires oxygen. This generates much more ATP than the anaerobic mechanism does. The pathways in the catabolism of fats and proteins, as well as carbohydrates, converge on a second key compound, acetyl-coenzyme A, which was discovered by Lipmann in the late 1940s. It is the acetyl group of this compound that is converted into carbon dioxide and water, via the citric acid cycle discovered by Krebs in the late 1930s. Utilizable energy is set free in various stages of these reactions, and transformed into a special kind of chemical energy, stored effectively in the pyrophosphate bonds of ATP.

This concept of the role of ATP, which makes it the central compound in energy transformations, was set out by Lipmann in 1941, and has been of great importance. It contains the essence of the answer to the question that was raised earlier: how does the food that we eat give us the energy that we need? The 'energy-rich' pyrophosphate bonds of ATP can be likened to a traveller's cheque in an universally acceptable currency. This form of energy is used to do mechanical work in muscle, to pump ions across membranes, and to build up macromolecules in the processes of biosynthesis.

The diversion of metabolism into catabolism and anabolism is not altogether clear-cut. There is a central area of metabolism where the pathways mingle. Nevertheless the routes for the synthesis of carbohydrates, fats and proteins differ from the routes for breakdown. The degradative routes were those first studied, and it was at one time thought that reversal of these routes might represent the synthetic pathways. But the actual arrangement, with separate routes, allows for a much greater degree of regulation than would otherwise be possible.

Exploration of cellular reactions has been accompanied by studies of cell morphology (Chapter 10); and much has been learned about the relationship between these two aspects. Study of an enzyme now usually includes an investigation of its behaviour during fractionation of the subcellular components; the results provide information about the intracellular localization of the enzyme. Certain types of reaction are associated with certain sub-cellular bodies or organelles. For example, the reactions of oxygen that lead to formation of ATP take place in mitochondria. Furthermore, when enzymes or enzyme systems are localized in organelles it becomes important to find out whether substrates and coenzymes can readily penetrate the organelles.

The detailed knowledge of metabolism provides a basis for certain generalizations about materials and processes. One example is the importance of phosphorus. The first clue was the observation of Harden and Young in 1905 that fermentation in the expressed juice from yeast quite soon came to a halt, but started up again when phosphate ions were added. Now there are many examples known of the natural occurrence of esters, anhydrides, and amides of phosphoric acid and two (ATP and thiamine pyrophosphate) have been mentioned above. In ATP the pyrophosphate groups take part in the metabolic reactions and the rest of the molecule plays only a supporting role. On the other hand, in thiamine pyrophosphate it is the thiamine part that is engaged in reaction and the pyrophosphate part that plays a supporting role.

Sulphur, like phosphorus, is an 'active' rather than a 'bulk' element; coenzyme A mentioned above contains sulphur. However, sulphur and phosphorus behave very differently. Group transfer (e.g. of methyl or acyl groups) takes place via sulphur which forms S—H and S—C bonds, and S—S bonds in redox reactions. But phosphorus does not commonly form such bonds and is seldom encountered except in the fully oxidised state. More generally, only a restricted range of elements are found in organisms, and there is a strong bias in favour of the lighter elements.

Similarly, molecules, or parts of molecules, keep cropping up in different contexts. Adenosine, for example, occurs in ribonucleic acids, ATP, coenzyme A, and the dehydrogenase coenzymes. Turning from substances to processes, we see that the route taken in the synthesis of the purine (adenine) in adenosine is exploited as an excretory pathway in the nitrogen metabolism of birds and reptiles. Despite the enormous variety of compounds in cells, there is a striking order and simplicity in the common underlying pattern of events. This has led to the idea that all organisms are derived from a common ancestor. It appears that life (as represented by the organisms now encountered) may have arisen but once. Studies bearing on the origin of life represent one of the newer branches of biochemistry. These studies are based on the premise that life arose in an anaerobic environment; it has been shown that a wide range of complex organic compounds can be formed from simple precursors (e.g. HCN) under these conditions.

We have stressed earlier that catalysis is a key concept in twentieth-century biochemistry. Cycles (e.g. the citric acid cycle) are an important type of catalytic system. Cycles can be regarded as a special kind of multi-enzyme system in which small molecules as well as enzymes are conserved components and play a catalytic role. The rate at which a cycle operates will depend on the activity and concentration of the enzymes, and the availability and concentration of the substrates. All these factors may be utilized in the regulation and control of metabolism. There are several differences between control of the concentration of an enzyme and control of its activity. In bacteria, the concentration of an enzyme may be raised when the enzyme is in demand for the metabolism of an unusual nutrient in the growth medium. On the other hand, an enzyme that catalyses one of the reactions leading to the synthesis of a compound may be repressible; the concentration of the enzyme drops when the compound is present in the growth medium. Enzyme formation here, is essentially *de-repression*. In animal cells, hormones that stimulate growth or development, may do so by affecting the rate of protein (and enzyme) synthesis. The activity of enzymes may be either increased or decreased. Decrease of the activity is seen in the *end-product inhibition* of metabolic pathways in bacteria; animal cells, on the other hand, generally display control by activation of enzymes. Regulation and control are impressively evident in differentiation, but the biochemistry of differentiation is as yet in its infancy.

The concept of information is helpful in discussing control. The term *information* is customarily used in the phrase ‘transfer of information’. The emphasis is on the direction, rather than the magnitude, of the information transferred. The transfer of information from DNA to protein† means that it is the kind of DNA present which determines what kinds of protein are made. The information is precise in just these circumstances where precision would ordinarily seem hard to attain. Proteins, as large and complex molecules, would seem intrinsically likely to be made incorrectly, but cells turn out many perfect copies. Thus the essence of the concept of the transfer of information lies in attempts to account for the regular happening of improbable events. Moreover, as is well known, the one-dimensional information in DNA is expressed in the three-dimensional conformation of protein. In this connection, Zuckerkandl and Pauling have written: ‘Here the concept of information replaces to advantage the concept of cause, since it is a better tool in the analysis of reality’.

Theoretical concepts in the biological sciences can be divided into two broad categories (Krebs 1966). One category contains those concepts that can be regarded as models (for example the energy-rich phosphate

† See Chapter 7, p. 162.

bond). The other, more general, kind of concept (e.g. evolution) does not arise as directly from experimental observations, but results from the urge to perceive order underlying diversity. Such general concepts give rise to subsidiary ones, such as the non-survival of nonfunctional characters. It seems certain that further general ideas will emerge, and we may hope that future historians of science will be able to take the second half of the twentieth century as the period in which theoretical concepts in biology started to flower richly.

Selected References

DIXON, M. and WEBB, E. C. (1964) *Enzymes*, 2nd edn. Longmans, London.

KEILIN, D. (1966) *History of cell respiration and cytochrome*. Cambridge University Press. (The topics discussed in this book have been omitted from my account of biochemistry.)

KREBS, H. A. (1966) in *Current aspects of biochemical energetics* (edited by N. O. Kaplan and E. P. Kennedy). Academic Press, New York.

ZUCKERKANDL, E. and PAULING, L. (1965) in *Evolving genes and proteins* (edited by V. Bryson and H. J. Vogel). Academic Press, New York.

7 Molecular Biology

The twentieth century is usually considered to be an age of increasing specialization with reduced contact between specialists in different fields. Rather curiously this generalization does not apply to biology. The advances of the last twenty years have drawn together previously disparate fields and there is today much common language between, say, a virologist interested in the viral infectious process, a biochemist studying the mechanism of protein synthesis, and a geneticist specializing in abnormal fungi. In view of this overlap between different fields, it is pointless to try to define too precisely the field of 'molecular biology'. This new term arose initially as a result of the contribution to biology of the physical technique of X-ray diffraction analysis, and now covers studies of the structure of large molecules and their interactions. Molecular biology is, of course, closely related to biochemistry, but it is really a broader term, covering the application of genetical and physical as well as chemical techniques to the study of biomolecular structure and function. To set the scene for discussion of these topics we must first consider briefly the main features of biochemical work since the turn of the century (see also Chapter 6).

It was in 1897 that the Buchner brothers first showed that a cell-free yeast extract could be prepared capable of catalysing the fermentation of sugar to alcohol. Since that time biochemists have isolated from cells and tissues a wide variety of specific catalysts, called enzymes, which bring about the many chemical transformations from one type of molecule to another, causing breakdown and synthesis, which are characteristic of the activity of living cells. Many of these enzymes have been purified and crystallized and they are found to be proteins—large molecules made up of amino acids strung together in peptide chains. Apart from their role as enzymes, proteins are also found as structural threads and building blocks: collagen in tendons; keratin in hair, nails, and feathers; fibrin in blood clots. Sometimes the two roles are combined, as in the threads of muscle cells, which are drawn past one another during contraction as a result of the enzymic action of the protein myosin which forms part of the thread structure.

The chemical and, to a large extent, the physical activity of a cell is controlled by enzymes, which are proteins; but what controls the enzymes? In broad terms this question was answered during the 1940s, when geneticists began serious study of the biochemical effects of mutations: enzymes are controlled by genes. We now know, as a result of recent advances, the chemical nature and structure of genes—they are long-chain nucleic acid molecules. A review of advances in molecular biology over recent years is essentially an account of how we have come to know so much more about protein and nucleic acid molecules, how these long-chain molecules are made, how the chains fold up and how they act. Specific genes determine the synthesis of specific proteins, by a process involving interactions between proteins and nucleic acids. This process is so central in biology that the new knowledge has had profound and far-reaching consequences in almost every field. Much remains to be discovered, for growth and differentiation depend on subtle interaction between the protein-synthesis system and the cellular environment which is still little understood. But our new understanding of the molecular detail of protein synthesis has brought the study of, for example, evolutionary processes, cell differentiation, virus infection, and antibody formation to a new level of sophistication.

It will be clear from what has been said already that this new knowledge did not arise *de novo*, or entirely from the imagination of the molecular biologists (though a surprising amount did arise in the latter way, as we shall see). The new advances depended primarily on the exploratory work of the early biochemists and geneticists: the purification of enzymes and their characterization as proteins (around 1930), chemical analysis and characterization of the properties of proteins, the main discoveries of classical genetics, the cytological demonstration that genes were situated on chromosomes (around 1920), the characterization of chromosomes as nucleoproteins (i.e. proteins complexed with nucleic acids†), and chemical analysis and characterization of the nucleic acids.

† See Chapter 8, p. 175.

Meanwhile (1910 to 1930), in physical study of crystals, X-ray diffraction analysis was proving an extremely powerful tool for the study of the structure and dimensions of small molecules.† During the 1930s Astbury, Pauling, and Perutz were considering in different ways how to bring such analysis to bear on the more complex problem of the structure of the large biologically-important protein molecules. Astbury was mainly interested in the fibrous proteins such as keratin, fibrin, and myosin. Placed in an X-ray beam, fibres give a relatively simple diffraction pattern of spots and arcs due to scattering of the X-rays at certain angles. These angles are directly related to repeat distances in the protein structure. For most of the fibrous proteins the molecular pattern along the peptide chain shows a repeat of 5–5·5 Ångstroms (1Å = 10^{-8} cm). Astbury (1935–40) suggested certain ways in which the molecular chain might be folded to give this repeat, and in 1950 Bragg, Kendrew, and Perutz published a paper in which they explored a large number of different types of folding, selecting eight for detailed study. None of these folds has proved to be the one mainly occurring in proteins. At this point the fundamental features of peptide folding remained elusive. In 1951, however, Pauling and Corey published a series of papers which threw new light on the problem. Their work, based on X-ray diffraction studies of crystals of amino acids and related small molecules, extended back to 1937. They determined very accurately from this preliminary work the lengths of the chemical bonds in a peptide chain, the angles between bonds, and the freedom of movement of molecular groups. From this data they made accurate molecular models and, without pre-conceived ideas on the subject, investigated the folds which these models could form. The spiral of Fig. 7.1, which they called the α-helix, came out as one of the simplest of a number of possible folds that satisfied the new rigorous atomic-relationship conditions built into the models. This helix has now been shown to be the type of folding most readily and commonly adopted by peptide chains.

We may ask at this point how Bragg and his co-workers, who were highly skilled crystallographers (and were soon after to unravel so triumphantly the detailed structure of the proteins myoglobin and haemoglobin), failed to consider the α-helix in their systematic study of possible peptide folds in 1950? The answer to this question is that in the crystals of small molecules, and for the packing of the protein molecules themselves in a protein crystal, only a restricted number of strictly symmetrical molecular arrangements are possible, since the crystal structure must repeat regularly in three dimensions. Crystallographers

† See Chapter 3, p. 82.

had been used to thinking in terms of 2-fold, 6-fold, and other regular symmetry axes and thought at first only in terms of folding and spiralling of peptide chains with an integral number of amino acids per turn of the spiral. All the helical folds considered by Bragg *et al.* in 1950 are of this kind. However, lengths of peptide chain within a protein molecule are not restricted by the symmetry considerations that govern the arrangement of the protein molecules as a whole in a protein crystal, and for peptide folding non-integral helices are possible. The α-helix,

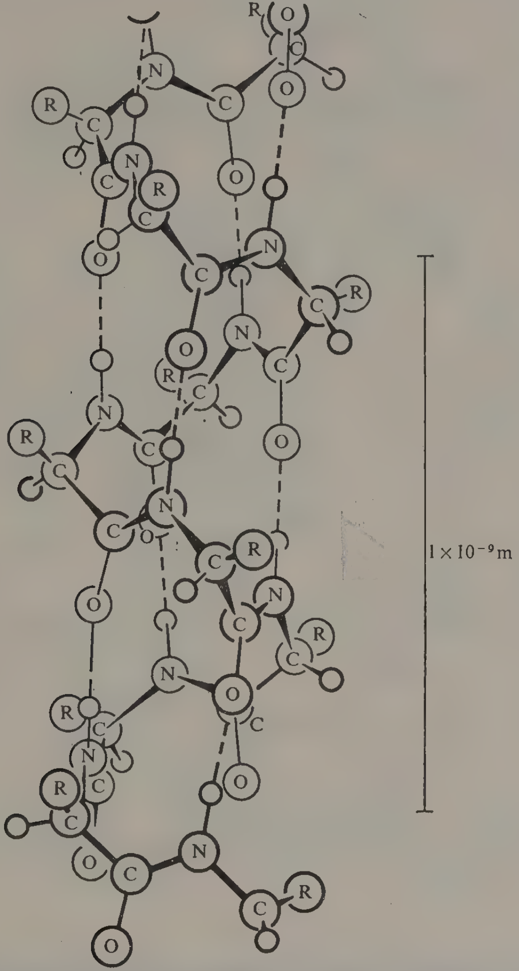


FIG. 7.1. The α-helix; C, carbon atoms; O, oxygen atoms; N, nitrogen atoms; ——o, hydrogen atoms; R, amino acid side-chains that are different for the different amino acids (from Pauling, L., Corey, R. B., and Branson, H. R., *Proc. natn. Acad. Sci. U.S.A.* **37**, 207 (1951)).

although unexpected, is in fact theoretically satisfactory and aesthetically pleasing, for in a helical structure, whether it has an integral or non-integral number of units per turn, it is possible for each repeating unit to stand in constant spatial relation to its neighbours. In the α-helix the exact spatial relation is determined by the hydrogen bonds and other interactions between the successive amino acids of the structure. Helices of one kind or another have now proved of fairly common occurrence in macromolecular structures, in nucleic acids as well as proteins.

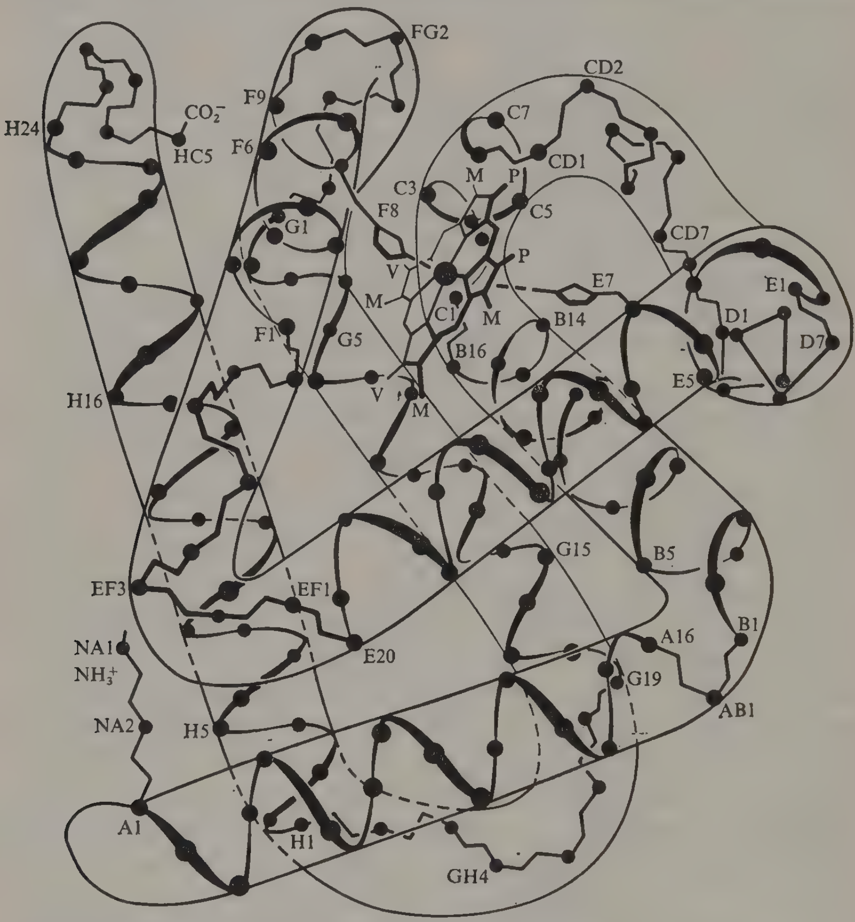


Fig. 7.2. Folding of the myoglobin chain. The letters NA1, etc. are indentifying labels that have been assigned to each amino acid of the sequence for convenient description and discussion. The haem group is seen at upper right with the peptide chain folding around it (from Dickerson, R. E. in *The proteins* (edited by H. Neurath), 2nd edn, Vol. 2. Academic Press, New York (1964).

We shall come to discussion of nucleic acid structure in due course but for the moment let us continue the protein story. At the time that Astbury was making his early studies of fibrous proteins, and Pauling and Corey of amino acid crystals, it was already known that crystals of globular proteins gave most complex and detailed diffraction patterns, full of information if only a way could be found to analyse and decipher patterns of such complexity. Around 1936, following the pioneer work of Bernal, Perutz set out to study the crystals of the oxygen-carrying protein of red blood cells, haemoglobin. For many years the problem proved intractable, and Perutz was meanwhile joined by Kendrew, who worked on the closely related but smaller protein myoglobin—the oxygen-carrier protein of muscle. The analyses for these two proteins were eventually carried through successfully by incorporation of metal atoms at specific points in the crystal, bound to the surface of the protein molecules. These atoms were located first and then used as markers in the analysis of the X-ray scattering from the protein molecules themselves. By 1960 the resolution attained was sufficient to show the complete folding of the

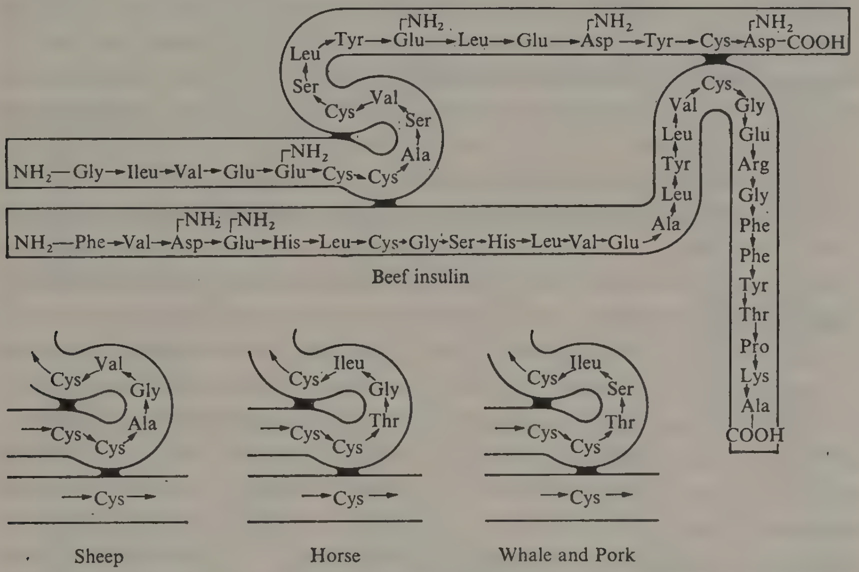

NH₂—Gly→Ileu→Val→Glu→Glu→Cys→Cys→Ala→Ser→Val→Cys→Ser→Leu→Tyr→Glu→Leu→Glu→Asp→Tyr→Cys→Asp–COOH
NH₂ NH₂ NH₂ NH₂
NH₂—Phe→Val→Asp→Glu→His→Leu→Cys→Gly→Ser→His→Leu→Val→Glu→Ala→Leu→Tyr→Leu→Val→Cys→Gly→Glu→Arg→Gly→Phe→Phe→Tyr→Thr→Pro→Lys→Ala COOH
NH₂ NH₂
Beef insulin
Cys Val Gly Ala Cys Cys
Cys
Sheep
Cys Ileu Gly Thr Cys Cys
Cys
Horse
Cys Ileu Ser Thr Cys Cys
Cys
Whale and Pork

FIG. 7.3. Primary structure of insulins of different species; *Gly*, *Ileu*, *Val*, etc. are abbreviations for the names of the amino acids; glycine, isoleucine, valine, etc. Cystines link chain points together through a chemical sulphur–sulphur bond) from *Introduction to molecular biology* (edited by G. H. Haggis). Longmans, London (1964).

chain in myoglobin (Fig. 7.2). The haemoglobin molecule was found to contain four peptide chains, two identical α-chains and two identical β-chains, all of about the same length and folded in much the same way as the myoglobin chain.

Meanwhile considerable progress had been made in the straight chemical study of proteins. Sanger, around 1950, devised techniques for determining the sequences of amino acids in peptide chains. This, like the application of X-ray diffraction studies to proteins, was a technical development of the greatest importance. Previously a balance sheet had been drawn out for a number of proteins, so many amino acids of one kind per molecule of the protein, so many of another kind: for insulin, for example, four valines, two isoleucines, and so on. Sanger showed that in the chains of the insulin molecule these amino acids were always joined together in the precise sequences shown in Fig. 7.3.

Sequences are now known for the haemoglobin chains, for myoglobin, and also for a number of other proteins, including the enzymes lysozyme and ribonuclease. For lysozyme, X-ray diffraction analysis has also been carried to the point where the three-dimensional folding of the chain is known. Taken together, the techniques for sequence determination and X-ray analysis have transformed our picture of protein molecules from vague blobs and fibres to precise structures with atoms in precisely defined positions. Questions such as how the haem groups interact to give the characteristic haemoglobin oxygen uptake curve can now be tackled at an altogether more precise and meaningful level, and, most exciting of current developments, Phillips and his co-workers are now studying the details of the binding of the substrate to lysozyme and have proposed a convincing scheme for the mode of action of this enzyme in which a number of amino acids around the binding site interact in subtle ways to bring about the splitting of the substrate molecule. As comparable structural studies are carried through to muscle proteins, membrane proteins, and so on, it will become possible to bring other topics of cellular activity under similar close scrutiny.

Even for the proteins we know already, elucidation of structure at this detailed level has produced further dramatic results. Sanger found that the insulin from a given species always had precise chain sequences, but in other species there were usually small differences, one amino acid substituting for another at one or other point in the chain (Fig. 7.3). The world-wide distribution of clinical laboratories equipped for simple blood assays and haemoglobin studies have made possible extensive study of haemoglobin from this point of view. Here it is not a question of species differences only, but of variation among individuals within the human species. The first abnormal haemoglobin to be studied in detail was sickle-cell haemoglobin. The sickle-cell molecule differs from the normal in only two amino acids; a glutamic acid at the same point in each of the β-chains is replaced by valine. This affects two parts of the surface of the molecule and makes it less soluble at low oxygen tensions. The red cells take up a sickle shape, get tangled together, and block fine blood vessels, causing further lowering of oxygen tension and further sickling. The sickle-cell condition is inherited in a Mendelian way and for someone with two sickle-type genes the condition is severe and fatal in early life. For someone who has inherited a sickle-type gene from one parent and a normal haemoglobin β-chain gene from the other the condition is not so severe. Part of his haemoglobin is normal and only part carries valine in place of glutamic acid.

Amino-acid sequences are controlled by genes—and this brings us to discussion of nucleic acids, and the experiments that have shown the chemical nature and mode of action of the genes.

It was clear by 1950, when Pauling and Corey were coming upon the non-integral spiral structure of peptide chains, and Sanger was working out the amino-acid sequence for insulin, that nucleic acids were worth studying as well as proteins. After all, the chromosomes that embodied the genes were made up of protein and nucleic acid, and even if the nucleic acid were just a scaffolding for genetic protein its structure might still be worth studying. So X-ray crystallographers, Astbury in the 1930s and more recently Wilkins and his co-workers, studied fibrils drawn out from viscous solutions of nucleic acids. Maintained at suitable humidity such fibrils give quite detailed X-ray diffraction patterns as a result of the spontaneous formation of oriented crystalline regions during the drawing of the fibrils. Analysis of these patterns was aided by theoretical study of the sort of patterns that could be produced by helices; this study was stimulated by Pauling's work on the α-helix. It began to look as though deoxyribonucleic acid (DNA) was also a helical molecule with two nucleic acid chains twisted together.

Watson and Crick's contribution at this point is now widely known. Nucleic acid chains are made up of linked nucleotides and in DNA there are four kinds of nucleotide: adenine, guanine, thymine, and cytosine. Watson and Crick proposed a detailed model for the DNA molecule inspired by the X-ray evidence for a helical structure and by chemical evidence that in DNA adenine (A) is present in the same proportion as thymine (T) and guanine (G) in the same proportion as cytosine (C).

In their model the arrangement of the nucleotides is such that A in one chain always pairs with T in the other, and G in one with C in the other:

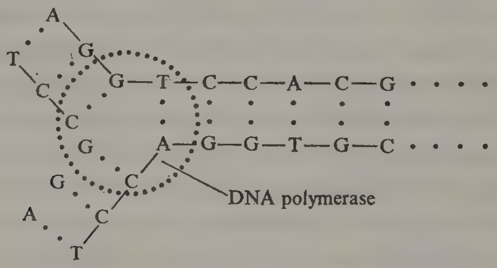
A—G—G—T—C—C—A—C—G ......

:   :   :   :   :   :   :   :   :

T—C—C—A—G—G—T—G—C ......

This is a structure in which there is no restriction of the order of nucleotides in one chain. In writing out the top line any of the four letters can be taken at random and set down in any order. But once the top line is written the bottom line is determined; the sequence of the second chain is complementary to that of the first.

Watson and Crick realized, in proposing this structure for DNA, that it might provide a molecular mechanism for the duplication of genetic material at each cell-division cycle. An enzyme might move down a double-stranded DNA molecule in such a way that the two chains came apart at the enzyme, with complementary nucleotides coming in and being joined to form new chains:



After the enzyme had moved right down the molecule two identical DNA molecules would now exist where one existed before. If the component of chromosome structure that carried genetic information were DNA, and if DNA duplication took place before each cell division, the daughter cells could each contain the same DNA as the parent cell. The nucleotide sequence along the DNA could then be thought of as a *coded message* carrying the inherited information. The mechanism for DNA duplication might usually work faultlessly, but occasionally, as a result of stray chemical and physical factors, a mistake in copying might be made. This would represent a mutation, and the altered DNA would be copied in subsequent generations. All this was implicit, and to some extent explicit as hypothesis, in Watson and Crick's first papers (1953).

These proposals both for the structure of DNA and the mode of its duplication may now be regarded as established by later more refined X-ray analysis of the DNA structure and by the isolation of enzymes

(DNA polymerases) that catalyse DNA duplication *in vitro*. The role of DNA as a carrier of genetic information may also now be regarded as firmly established as a result of experiments to be described that have elucidated the mechanism of protein synthesis. These experiments have shown that the DNA message is a code specifying amino-acid sequences in protein peptide chains. A given length of DNA (one gene) specifies the amino-acid sequence for one specific peptide chain. If one of the nucleotides is altered, then, according to the position of the alteration, the change will in some cases not alter the corresponding amino acid, while in other cases a different amino acid will now be found at the corresponding point in the peptide chain. At some point, perhaps long back in time, an ancestor of a person now suffering from sickle-cell anaemia was subject to a mutation at one point on his or her haemoglobin β-chain gene. The DNA inherited from this ancestor now codes at this point for valine instead of glutamic acid.

Before we discuss protein synthesis it will be useful first to describe briefly the development of the electron microscope, which has enabled us to visualize the cytoplasmic particles where protein synthesis occurs, and other aspects of cell ultrastructure. In certain cases, notably in the study of viruses and the contractile apparatus of muscle cells, the electron microscope has made a contribution to our understanding of structure and function at the molecular level as important as the contribution of X-ray diffraction. More generally, in the study of cell ultrastructure, the electron microscope brings morphological study down to the near-molecular level, and shows the framework within which molecular events are operative in the cell. Electron microscopy and X-ray diffraction studies are thus complementary, the one taking over where the other fails, with some overlap in favourable cases.

The electron microscope was developed by Knoll and Ruska in the 1930s, growing out of their work on electron beams. Electrons travel in straight paths in a vacuum but can be deflected by electric and magnetic fields. Thus suitably designed electromagnets can be used as electron ‘lenses’. The arrangement of these lenses in an electron microscope is very similar to the arrangement of the lenses of a light microscope, except that the specimen and lenses of an electron microscope have to be placed in an evacuated column (for unless the components are in a vacuum the electron beam is scattered by collisions with gas molecules). The image has to be formed by the electrons hitting a fluorescent screen or a photographic plate. Further, the specimen in an electron microscope has to be very thin, since electrons have poor penetration.

It might have been expected that a microscope using electrons instead of light waves, with very different wavelength and scattering properties, would have required radically different methods of specimen preparation. At first this was true. In the early biological use of the electron microscope a method called 'shadowing' was developed for looking at virus particles in which the particles were coated with metal on one side in a metal evaporation unit. The electron scattering power of the metal provided the necessary contrast, and the final print could be made with the metal light and the uncoated side of the particle dark, so that the virus appeared to be lighted from one side, throwing a dark shadow. Later, in the 1950s, methods were developed for looking at cells and tissues, and the most successful of these have proved to be simple extensions of the methods used to prepare tissue sections for light microscopy—simple extensions in principle, that is, but technically very demanding for the pioneers. The fixative or stain has to contain a metal to give electron scattering and contrast, the embedding medium has to be plastic rather than wax to get very thin sections, and high-precision microtomes have to be used.

These developments have opened up a new dimension in cytology. The mitochondria of the cell cytoplasm, which are the power generators of the cell, mobilizing metabolic energy in the form of adenosine triphosphate (ATP)† are seen simply as small granules and rods in the light microscope. In the electron microscope they are seen to be composed of an outer membrane and an inner membrane, the latter thrown into folds called 'cristae'. Bound to the cristae, and partially visible as knobs 90Å in diameter, are the enzymes of the phosphorylation reactions which form ATP. Here the electron microscope is probably showing enzymes or groups of enzymes, but cannot reveal their detailed structure or mode of action.

In the muscle cell the electron microscope shows two sets of filaments that slide past one another during contraction: movement cannot be seen unfortunately in the fixed electron microscope preparations, but this sliding movement is deduced from the increased overlap between the filaments in muscle fixed after contraction. One set of filaments is made up predominantly of the protein myosin and the other of the protein actin. Protrusions can be seen on the myosin filaments, which are probably the sites at which the two filaments interact. These protrusions carry the sites at which ATP is split to provide the energy for contraction. Again we can see an enzyme in the electron microscope in a position of

† See Chapter 6, p. 150.

obvious functional importance, but cannot see its exact structure or mode of action.

We can return now to the protein synthesis story. In the link between DNA and protein synthesis the main role is played by a different type of nucleic acid closely related to DNA. This is ribonucleic acid (RNA). In cells that are synthesizing a lot of protein, as was shown around 1940 by Caspersson and Brachet, there is a relatively large amount of RNA present in the cell cytoplasm. Caspersson showed this from the ultraviolet absorption due to the nucleic acid, and Brachet by specific staining, and enzyme studies, of light-microscope preparations. In electron-microscope sections, in the cytoplasm of cells active in protein synthesis many particles about 250Å in diameter are seen, some free, some attached to cytoplasmic membranes. These particles, and fragments of cytoplasmic membranes with attached particles, can be recognized by electron microscopy in fractions of homogenized cells separated by differential centrifugation. From chemical study of the fractions the particles can be shown to be made up of RNA and protein, and are termed *ribosomes*.

Studies with amino acids labelled with a radioactive isotope of carbon show further that ribosomes are directly involved in protein synthesis. In early experiments of this kind carried out around 1955 labelled amino acids were injected into the circulation of an experimental animal; pieces of tissue such as liver were then homogenized at varying times after injection of the label, and separated into sub-cellular fractions by differential centrifugation (that is, by spinning at successively increasing speeds). When the cells were homogenized a short time after injection of the labelled amino acid most of the radioactivity was found to be associated with the ribosomes, but if a longer time was allowed before homogenization it was found associated predominantly with soluble protein. This showed that peptide chains in course of synthesis must be bound to ribosomes.

Further important advances were made by Zamecnik and Hoagland during 1955–6. They found that in the first step of protein synthesis, catalysed by specific enzymes, the amino acids interact with ATP. The ATP is split and this is the energy-supplying step for the eventual assembly of the amino acids into chains; these enzymes are therefore termed 'activating' enzymes. The amino acid is then linked to an RNA molecule. The RNA molecules involved at this point differ from the ribosomal RNA in being of much lower molecular weight. In fact they were at first called *soluble RNA molecules*, because they remained in the supernatant solution when the ribosomes were spun down in a centrifuge. Later they have usually been called *transfer-RNA molecules*, for it has been found that their role is to carry the amino acids to the ribosomes for assembly into peptide chains. It has been found that there are at least twenty different kinds of transfer-RNA molecules corresponding to the twenty main amino acids required for protein synthesis, and twenty different activating enzymes linking each amino acid specifically to its own kind of transfer-RNA.

This seems, at first sight, a rather complicated biochemical mechanism for joining amino acids together, and before carrying the story further it will perhaps be useful to take a look at protein synthesis from a more theoretical and less biochemical point of view. There was already by 1953 much genetic evidence to suggest that genes exert their control over cellular activity mainly through control of protein synthesis. From the time of Watson and Crick's proposal for the structure of the DNA molecule, molecular biologists, particularly Crick himself, began to speculate about the way a molecule of this kind might exert this control. The central idea underlying this speculation was that the nucleotide sequence in the DNA must be in some way a code message specifying amino-acid sequences in peptide chains. If a DNA code message using only four types of nucleotides is to specify sequences in peptide chains made up of twenty different types of amino acid, the code cannot be achieved by one-to-one correspondence between a nucleotide and an amino acid. Nor can the code be two-to-one since four types of nucleotide taken in pairs can form only sixteen combinations. Conceivably four of the amino acids might be specified one-to-one by the four bases and the other sixteen amino acids by the sixteen double combinations. But Crick's hunch, from about 1955, was that it was probably more feasible at a molecular level to envisage a triplet code with each successive sequence of three nucleotides along the DNA coding for successive amino acids of the corresponding peptide chain. This hunch was justified in due course by genetic experiments with triple mutants, which will be described later.

Crick further suggested that there must be some amino-acid 'adaptor' mechanism similar to the transfer-RNA mechanism later discovered. Nucleotide triplets do not really have the right sort of molecular groups to bind specifically to amino acids, but it is clear that in the DNA structure a given triplet can bind specifically to the complementary triplet. Crick suggested therefore that during protein synthesis amino acids must be first attached to adaptor molecules carrying nucleotide triplets. The triplets of the adaptors could then bind to successive triplets along a template nucleic acid, hence bringing amino acids into a sequence determined by the template nucleotide sequence. This is in fact what is found. The activated amino acids are linked to transfer-RNA molecules, and one triplet of the transfer-RNA sequence (the anticodon) is complementary to the coding triplet (codon) for the corresponding amino acid.

It was thought probable, around 1960, that the sequence of nucleotides in the DNA of the genes controlled the synthesis of ribosomal RNA, the ribosomal RNA being synthesized by a mechanism very similar to DNA duplication, so that the ribosomal RNA carried the gene's code message. Then as transfer-RNA molecules came up to the ribosomes it was supposed that codons and anticodons came together so that amino acids were aligned and subsequently joined in a sequence determined by the sequence of codons in the ribosomal RNA. This scheme had to be modified however as a result of the discovery in 1961 of a new kind of RNA molecule intimately involved in protein synthesis—a messenger-RNA carrying code message, from the genes to the ribosomes, which differs from the main ribosomal RNA. This messenger-RNA was first studied in phage-infected bacteria in which the protein-synthesis machinery goes over, after infection, to synthesis of phage protein. It was found by isotope-labelling experiments that this new protein was synthesized on the old ribosomes present before infection, under the influence of RNA formed under control of the phage DNA. Evidence was soon found for a similar type of messenger-RNA in uninfected bacteria, and later RNA molecules with similar properties were found in plant and animal cells.

It was concluded, therefore, that the bulk RNA of the ribosomes must play a purely structural role, and a picture of protein synthesis emerged that is now thought to be substantially correct. Genes concerned with specification of proteins are thought to specify sequences for the appropriate messenger RNA molecules, synthesized by RNA polymerases, acting like the DNA polymerases of DNA duplication but forming RNA molecules complementary to only one strand of the DNA. These messenger RNA molecules move, in general, from the nucleus (where the DNA is) to the cytoplasm (where most of the ribosomes are), and the ribosomes bind to and move along the messenger. As each messenger codon comes opposite the appropriate point on the ribosome the appropriate transfer-RNA molecule is bound to the ribosome and the amino acid it brings is joined to the growing peptide chain. The ribosomal structural RNA and also the transfer RNAs are thought to be synthesized by segments of DNA devoted to this specific purpose.

Thus, within less than a decade of Watson and Crick's proposal for the DNA structure, the main features of the mechanism of genetic control of protein synthesis had been elucidated. But there have also been important advances since 1961, notably in deciphering the DNA code. First, Crick and co-workers provided important genetic evidence confirming the theoretical hunch favouring a triplet code, and showing that the successive codons for a given peptide chain were read sequentially from a certain starting-point on the DNA. They studied phage mutants in which there had been addition or deletion of a base in the DNA sequence (rather than an alteration of a base, which is the action of some chemical mutagens). They showed that insertion or deletion of one or two bases completely messed up the reading of the corresponding gene, since the message, read in triplets from one end, became completely jumbled. Insertion or deletion of three bases sufficiently close together in the sequence led, however, to a normal or nearly normal phenotype. Most of the message, beyond the insertions or deletions, now read as in the wild-type gene.

Also increased understanding of the mechanism of protein synthesis allowed development of *in vitro* systems containing ribosomes, activating enzymes, transfer RNAs, amino acids, ATP, and certain other necessary factors. These systems could be stimulated to synthesize protein by addition of messenger RNA, and it was found by Nirenberg and his co-workers that polynucleotides of restricted nucleotide composition could also act as messengers and produce peptide chains of restricted amino-acid composition. In the simplest cases, poly A produces a peptide chain containing only the amino acid lysine, poly C a chain containing only proline, etc., so that AAA codes for lysine and CCC for proline. From experiments with more complex polynucleotides and from study of the binding of trinucleotides to ribosomes the full code has now been determined. Some triplets act as chain-terminating signals, and a special class of transfer-RNA molecule is involved in chain-initiation. The *in vitro* peptide synthesis experiments, as well as elucidating details of the code, also provide very effective confirmation of the whole scheme of genetic control of protein synthesis.

Although we now understand in some detail how genes determine the proteins that can be made in a cell, major problems still remain unsolved in our understanding of the control of the genes, and in our understanding of the coordination of cellular activity. Genes are not altered by their environment, except by mutation, and this is generally a rare and random process; the genes present in an organism are determined only by parental inheritance. On the other hand, the activity of these genes in any given cell at any given moment is very much affected by environmental factors. As cells become differentiated during embryogenesis many of the genes are somehow rendered inoperative (although still present) and cease to produce the appropriate messenger RNA. In a differentiated cell only a restricted range of proteins are formed, with a few proteins predominating; for example, myosin and the other proteins of the contractile apparatus in muscle cells, haemoglobin in the developing red cell, and so on. The only real clue we have to the nature of these effects comes from the study of enzyme induction in bacteria. Bacteria can in some cases produce when required the enzyme they need to cope with a small molecule in their environment—to incorporate it if it is nutritious or modify it if it is toxic. To show this effect they must already carry the gene for the required enzyme in their inherited gene complement, but the gene can be inactive in a normal environment and switched on by the small molecule. It seems that certain proteins called *repressors* are present in bacteria, and presumably also in the cells of higher organisms, which can complex with DNA at certain specific points in the nucleotide sequence and specifically block the expression of some genes. Specific small molecules, taken up into the cell from the environment, can interact with these repressors and loosen their binding to the DNA. This idea is probably of general application, but provides only a basic principle. We still know very little, in molecular terms, of how a fertilized ovum develops into a formed adult organism.

In the field of coordinated cellular activity there is also a big gap between, on the one hand, the spontaneous folding of a peptide chain to form a globular unit or spontaneous formation of simple protein aggregate structures, which we can begin to understand now in molecular terms, and, on the other hand, the coordinate movement of chromosomes to the equatorial plane, and all the subsequent complex features of cell-division.

Even the spontaneous assembly of more complex cell structures is difficult to envisage, quite apart from the further problems of coordinated movement and activity in the cell. In muscle cells, for example, the myosin and actin molecules must aggregate to form the two types of filament. The myosin filaments then become lined up in hexagonal array interdigitating with the actin threads to form the muscle fibril. Mitochondria proliferate to fill the spaces between the fibrils. It is not known with any certainty how new mitochondria are formed, perhaps by growth and division of existing mitochondria. In developing mitochondria a whole range of enzymes and other proteins somehow assemble into specific complexes attached to the mitochondrial membranes. Some mechanism must exist controlling their specific attachment to the inner mitochondrial membranes and not to the other membranes of the cell cytoplasm.

In these assembly processes carried out by cells there are probably a number of factors involved. In the first place there may be control exerted by the time-sequence of synthesis of the relevant proteins. If, for example, two proteins are formed first and complex together, and then a third type of protein is synthesized, a final structure may be formed in a more controlled way than could be achieved by complexing all three proteins together at the same time. A second factor is the presence of preformed structures. The presence of certain lipid molecules, for example, in some of the cell membranes may lead to specific adsorption of certain enzymes that synthesize more of the same lipid. A further factor is the presence of some DNA in cytoplasmic organelles as well as in the nucleus. It has been known for some time that, in additon to chromosomal genes, extrachromosomal genes are present in many organisms and can be recognized and studied by genetic methods although they do not segregate in a Mendelian way. It is known that there is DNA in mitochondria as well as in plant chloroplasts, and some evidence from work on *Neurospora* suggesting that the mitochondrial DNA carries extrachromosomal genes, notably the gene coding for the structural protein of the inner mitochondrial membrane. The presence of DNA in cytoplasmic organelles such as mitochondria may be of importance in two ways: first, control of the duplication of this DNA can be separated from control of duplication of the cell as a whole, allowing separate control of mitochondrial proliferation. Second, synthesis of certain proteins coded by the mitochondrial DNA may take place only within the mitochondria. If these proteins are formed in the inner compartment of the mitochondrion and are of a type which tend to bind to membranes, then they will become bound to the inner surface of the inner mitochondrial membrane. Here they could form a surface specifically adsorbing other proteins of the mitochondrial enzyme complex.

Finally, to draw a general conclusion from all these developments that summarizes their impact on biology as a whole, one may say that the whole movement has been towards greater precision of thought. Genes are no longer thought of as beads on a string, or enzymes as globules of protinaceous matter. In the DNA structure the positions of the individual atoms are precisely determined, and in some cases, of which we may take sickle-cell anaemia as an example, we know that a mutation has altered thymine to adenine at a certain point in the DNA structure, and we know how this leads to vascular accidents and death.

(We do not know, unfortunately, how to set the altered DNA right again.) For one enzyme, lysozyme, the positions of the atoms are also known with precision and in this instance we know how an enzyme surface catalyses a chemical reaction—the basic chemical activity of life.

These advances are fundamental to the whole of biology and medicine, although their practical impact will probably not be immediate. One cannot yet claim that increased understanding of protein synthesis, or of the molecular events underlying mutations, has helped very much towards discovery of a cure for any specific disease, or of a new antibiotic, or a new fungus-resistant breed of wheat. There is always the possibility that a new fundamental advance will provide a dramatic cure for cancer, or alternatively as McFarlane Burnett fears, produce some fantastic calamity such as the production, in the conditions of a molecular biology laboratory, of an uncontrollable virus. But neither of these dramatic events is very probable. Increased fundamental understanding is more likely to have mainly a long-term effect in providing new tools, both experimental techniques and modes of thought, for the study of more practical problems. Clearly, however, this 'break-through' in biological science is going to give us much increased opportunity to tamper with and modify biological material, with possible beneficial or harmful social consequences.

Outside the purely scientific and social impact, and at a more philosophical level, the main effect of increased molecular understanding of living processes, with the methods and thought-processes of physicists and chemists penetrating further into biology, has been to blur and finally to obliterate the dividing-line between living and non-living. The new knowledge shows in what subtle ways the normal forces between atoms and small molecules, which operate in physics and chemistry, operate also between large and small molecules in the biological situation of the cell. A continuous evolutionary chain can now be traced, in principle, from atomic interactions, through macromolecular aggregates and primitive cells to cell interactions, multicellular organisms, nerve-cell contacts, memory mechanisms, and so to inherited and acquired behaviour-patterns.

There are still many gaps in our detailed understanding of this process, and it remains to be seen how far human behaviour in all its aspects can be described in these terms. But it seems that the ultimate achievement of the scientific method may be to show how human thought and personality arise through an almost unbelievable sequence of linked patterns of increasing complexity from the random movement of atoms and the laws of atomic interaction—the laws that govern the formation and interactions of the sub-atomic particles for which modern physics still seeks precise expression (see Chapter 3).

References

CAIRNS, J., STENT, G. S., and WATSON, J. D. (editors) (1966) *Phage and the origins of molecular biology*. Cold Spring Harbour Laboratory. A history of how work got started on the genetics of bacterial viruses (phages).

CRICK, F. H. C. (1966). The genetic code. *Scient. Am.* **55**, Oct.

JACOB, F. and MONOD, J. (1961) *J. molec. Biol.* **3**, 318. The authors propose a mechanism of gene regulation based on study of enzyme induction in bacteria.

PHILLIPS, D. C. (1966) The three-dimensional structure of an enzyme molecule. *Scient. Am.* **78**, Nov. X-ray diffraction study of lysozyme and a proposal for its detailed mode of action.

WATSON, J. D. (1968) *The double helix*. Weidenfeld and Nicholson. A history of the development of ideas on DNA structure as seen through the eyes of one of the workers mainly involved.

8 *Ecological Genetics*

In the early 1920s, zoological studies in most universities of the world were largely adjusted so as to display and develop the evidence for evolution, but not its mechanism. This, of course, was so at Oxford, where zoology consisted principally of comparative anatomy and embryology. In addition, however, and rather isolated from the main trends of study and research there, were some excellent lectures on genetics delivered by Sir Julian Huxley. But that subject had not then become evolutionary except in so far as it demonstrated the existence of variation upon which, clearly, selection could operate. In fact, seeing the Oxford course of zoology as an undergraduate at that period, I could not but feel how indirect and inferential was the evidence presented to us for the working of natural selection. Nor would this have been different elsewhere at that period.

It is true that here and there during the first quarter of this century a few attempts had been made to treat evolution experimentally. They were, however, isolated, while genetic knowledge was then insufficient to sustain them. Indeed, for the idea of a definite experimental approach to this subject, albeit one that never directly bore fruit, we should have to go back further. At some time in the mid-1920s, Major Leonard Darwin told me of a striking conversation remembered from his childhood. It was held with his father, Charles Darwin, who maintained that it should be possible to detect and analyse evolution actually taking place in nature today. For this, he said, a programme of observation and experiment would be necessary extending, so he thought, for perhaps fifty years in respect of forms reproducing annually. In view of some work of my own begun during the First World War, I was convinced that Darwin was right but that, in respect of the period of time needed to obtain results, he was much too pessimistic.

In 1927, moreover, Fisher published a paper on the theory of mimicry, in which he envisaged the selective modification of the *effects* of genes in natural conditions. He followed it the succeeding year with his theory of dominance-modification (Fisher 1928), which in fact involved another aspect of the same concept; one that, in its entirety, must represent a fundamental component in the evolution of wild populations. It appeared therefore that the time had arrived for developing a project to detect and analyse selection taking place in nature. Clearly it should employ the basic techniques of science: those of observation and experiment, so curiously lacking in evolutionary studies up to that period though evidently envisaged by Darwin himself.

It seemed to me that this object could be attained by a joint programme of laboratory genetics combined with observation and experiment in the field, using the statistics appropriate to such studies. These, moreover, had at that time recently been supplied by Fisher in his now famous book *Statistical methods for research workers*, first published in 1925.

This twofold technique is today known as ecological genetics (Ford 1965*a*, pp. 2–3). It is essentially evolutionary in aim, for it deals with the adjustments and adaptations of wild populations to their environment. Indeed, as I have stressed in the past, it provides the only direct means for studying and assessing the actual process of evolution taking place at the present time. On the other hand, ecological genetics is synonymous neither with evolutionary genetics nor with population genetics, both of which include laboratory studies and the development of mathematical theories quite divorced from ecology. Moreover, some aspects of ecological genetics are but indirectly related to evolution, nor are they concerned with population studies; as, for example, the genetic analysis of adaptations, and the comparison and assessment of apparently similar phases in distinct communities.

Yet, among the aims of ecological genetics, the study of evolution in wild populations is surely outstanding. However, it can only be undertaken efficiently when the changes involved occur rapidly. Three situations in which they do so were picked upon at the outset of this work, and time has shown that they have indeed supplied the conditions necessary for conducting it. That is to say, the effects of selection prove to be sufficiently powerful for effective study (1) when isolated populations undergo pronounced numerical fluctuations; (2) when polygenic characters (p. 177f) can be examined either in populations inhabiting distinct and isolated areas or else under powerful selection even in the absence of isolation; (3) in all instances of genetic polymorphism.

In order to discuss the way in which evolution can take place in these three situations, it is necessary to consider very briefly a few propositions fundamental to heredity. In the first place, it is clear that some sort of *units* responsible in some kind of way for the characteristics of each organism must be transmitted from parent to offspring if heredity has any physical basis at all. These units are known as *genes*. It can be shown by a simple mathematical device, that of 'correlation', that the two sexes contribute in equality to the heredity of their offspring. Consequently, the genes must on the simplest basis be present in pairs, the members of which, known as *alleles*, are derived respectively from the male and female parent.

The two genes constituting each pair must, however, separate or '*segregate*' from one another at the formation of the germ cells, which are contained in the pollen grains and ovules of plants and constitute the sperm and the eggs of animals; for only in that way can the heredity of the succeeding generation be bisexual also. Thus in their hereditary component the germ cells constitute half a cell each: the pairs of genes are restored once more at fertilization since this is an additive process.

The genes, though responsible for given sets of characteristics, can exist in two or more forms, producing variation in the features they control. For evidently the pairs of alleles may be *homozygous* made up of similar members that is to say, or they may be dissimilar (*heterozygous*).

Yet the genes are very constant, for they do not contaminate one another, or 'blend', when brought together into the same individual. They are, moreover, intrinsically stable, for they very seldom undergo chemical alterations, known as *mutations*: changes in the units of heredity, that is to say, giving rise to new 'mutant' genes and characteristics. Thus the genetic variability of organisms is due to the infinite opportunities for recombining the hereditary material in new ways provided by segregation, hardly at all to mutation. Such recombination, then, is the main source of genetic diversity.

The pairs of genes are carried in paired structures, the *chromosomes*,† the members of which (*homologous chromosomes*) also separate from one another at germ-cell formation, to be recombined again when fertilization occurs. Obviously they could not act as vehicles for the

† For the chemical aspect of chromosomes and genes see Chapter 7.

transmission of the hereditary units unless this were so. The chromosomes are few in number, twenty-three pairs in Man, so that each carries large numbers of genes, which are said to be 'linked' with one another. These then must be held together in groups each representing those located in a particular chromosome. The alleles comprising the respective pairs can be transferred in blocks from one homologous chromosome to another ('crossing-over') because a few interchanges of material take place between them at the period when segregation occurs.

Thus the hereditary material is equipped with two apparently inconsistent properties: great heritable variability, due to segregation and recombination, providing the diversity upon which natural selection can work; also great heritable stability, to preserve qualities, and combinations of qualities, that are advantageous to the organism.

Such stability would be impossible were the genes to contaminate one another, or were they chemically unstable: that is to say, subject to a high mutation-rate. If, then, mutations were common enough to produce the variability upon which selection can work, or to adjust the body directly in some way to fit it to its environment, the genetic material would be too impermanent to maintain the advantageous qualities upon which the organism depends.

We can therefore reject as impossible all methods of evolution that are supposed to work by controlling mutation. Of these Lamarckism, involving the inheritance of the effects of use and disuse, is one; another is Buffonism, which implies the direct inheritance of environmental effects.

We may now turn to a brief survey of the three situations in which evolution occurs rapidly enough in natural conditions for convenient study.

Numerical fluctuations in isolated communities

The heritable variability of organisms must be balanced between the effects of natural selection, tending to reduce it, and of genetic recombination tending to increase it. So too, the numerical expansion of a population must be limited by conditions that check the effects of potential reproductive capacity. Such limitations must bear unequally upon individuals of different types while, in so far as their basis is inherited, they will spread certain characteristics through the community and tend to destroy others: for some forms will be relatively well, and others ill, suited to the prevailing environment.

To put it the other way round, a population can increase in numbers only if certain aspects of selection bearing upon it are relaxed. In such circumstances it will therefore become more variable. On the other hand, a numerical decline indicates more rigorous selection, and this must tend to reduce diversity. It will be noticed that variability, being at random relative to the needs of the organism, will be much more often disadvantageous than advantageous, so that characteristics adversely affecting the adjustment of the individuals will tend to spread through a community when it is expanding. Subsequently, therefore, there will be more variation to eliminate when selection becomes stricter again, and in consequence the population declines numerically, than would have been present in a community that had not previously expanded. Thus numerical increase with relative diversity prepares the way for numerical decline with relative uniformity.

The cycle so generated is one that tends to promote rapid evolution. For, during the period of increase, genes can be tried out in a variety of fresh combinations that may occasionally prove advantageous to the organism. Yet such new assortments of the hereditary material might take an immense time to realize in a numerically constant community.

These simple conclusions had been reached when we were confronted with an isolated and rapidly expanding population of a butterfly, the Marsh Fritillary, *Euphydrias aurinia* (Ford and Ford 1930), during the period 1920 to 1924. This species has but one brood in the year and in the locality in question it had long been rare and very constant. Yet during the period of numerical increase the insects were affected by an extraordinary outburst of variation. It was very noticeable, moreover, that the more striking and exceptional forms tended to be deformed while some of the most extreme were actually unable to fly.

By 1925 the butterfly was very abundant, but from then on it became no commoner. Meanwhile its variability, so marked when the population was expanding, was checked and the locality came to support a relatively constant form, which, however, differed from the one that characterized it before the period of numerical increase. That is to say, the insect had indeed availed itself of the opportunity for rapid evolution provided by a marked fluctuation in numbers.

Evolution in polygenic characters

Variation is, in general, partly environmental and partly genetic in origin. When the latter component is ‘polygenic’ it is due to the action of a large number of genes having small cumulative effects. Thus the variability involved does not fall into distinct classes, as with the human blood groups (pp. 187-9), but is of the 'continuous' type illustrated by human height when examined in one sex of a population belonging to a single race. That situation is favourable for rapid evolution since the individual effects of segregation are then small, so that each does little to disturb the genetic balance of the organism. Moreover, the action of such selection is gradual, allowing opportunities for the progressive repairing of any harmful effects that may be produced. If the population involved is isolated, it can, as was clearly recognized by Darwin, be adjusted to changes in its habitat without dilution by individuals adapted to other conditions.

We have for many years studied adaptation and evolution in these and other circumstances in a butterfly, *Maniola jurtina*, Satyridae (Ford 1965*a*, Chapter 4). It feeds on grass, has one generation in the year and is extremely widespread in Britain. Our work upon it in the Isles of Scilly provides excellent illustrations of the conditions just mentioned. In this insect we required a feature that could be studied quantitatively. We therefore used the number of spots on the underside of the hind wings, which can take any value from 0 to 5 on each side. Their presence can be employed as a criterion of variability, by means of which it is possible to compare populations and examine changes in them.

Genetic work conducted in the laboratory by McWhirter (1968) has assessed heritability of spotting in this insect. As a result of recent studies based on larger numbers, it proves to be 0·63 ± 0·14 in the female: that is to say, approximately 60 per cent of the observed variation is hereditary and 40 per cent is environmental. In the conditions in which the breeding was carried out, the heritability is much smaller (0·14) and is indeed not significant in the males, though evidence is accumulating to show that this value may be greatly increased in other environments; thus at a high temperature (22°C) in the laboratory it reaches 0·4, while in females it approaches 1·0.

The islands upon which we have examined this species in Scilly fall into two classes: large and small, with areas of 682 acres or more compared with 40 acres or less; a difference of at least 17 times.

We may here limit ourselves to a consideration of the female spotting, since this differs qualitatively in the more striking way. It is identical, with approximately equal numbers of specimens having 0, 1, and 2 spots, on the three large islands on which we have studied it, while it differs consistently from one to another of five small islands; taking such values as unimodal at 0 or at 2 spots, or bimodal with the greater mode at either of them.

Various suggestions, including different aspects of chance survival or 'random genetic drift', have been advanced to explain the diversity of spotting from one to another small island compared with its similarity on all three of the larger ones.

Here we need consider only the explanation, based upon powerful selection, which has now been established. That is to say, organisms can be adjusted to the special ecology of small areas, provided they are isolated, but only to the average of large diversified ones; and averages tend towards uniformity. We should therefore expect populations to be more alike when we compare them on a number of large islands than when we do so on a number of small ones.

We have, indeed, already been able to study the reaction of *Maniola jurtina* to a marked ecological change occurring on a small island (Tean). This was due to the removal in the autumn of 1950 of a herd of cattle that had long been maintained there. It produced an alteration from lower to higher spotting, the respective types remaining constant each season before and after that event. This adjustment involved selection of 64 per cent (with 95 per cent probability limits at 83 and 27 per cent) against non-spotted individuals in a single generation. We found, that is to say, a genetic response to a modification in the habitat; for the absence of grazing and trampling by the cattle decisively altered the vegetation of the island. The resulting difference in spot-frequency was as great as that between one and another of the small islands in Scilly.

Other instances of the kind have now been witnessed. Attention may here be drawn to one of them on account of the extraordinarily high selection-pressures involved. White Island, another of the small islands of Scilly, consists of two high rocky areas connected by a low neck of land. In the winter of 1957–8 the sea washed across this isthmus, covering it with a belt of shingle about 30 metres wide and so converting it into a partial barrier to the butterfly. The insect had previously been of a similar type over the whole area, one with a heavy excess of spotless females. After the incursion of the sea, the spotting remained similar on the northern part of the habitat but changed markedly on the southern one, for a population having approximately equal numbers of females with 0, 1, and 2 spots established itself there and has remained of that type in succeeding years. The selection against the spotless females in this region amounted to 80 per cent.

It may seem surprising that a belt of shingle no more than 30 metres wide should constitute a partial barrier to this butterfly, a species quite powerful on the wing. Yet we find that it can build up distinct communities even without isolation, provided it is subject to the impact of strong selection.

Across southern England from the North Sea to Devon, and indeed throughout Europe from Bulgaria to the Pyrenees, the *Maniola jurtina* population is of the type formerly inhabiting the whole of White Island; that with a 'unimodal' excess of females at 0 spots. In east Cornwall and west Devon, however, this is replaced by a 'bimodal' form with two spot maxima in the females: a greater at 0 spots and a lesser at 2 spots. The one changes to the other in a few yards even though no physical barriers separate them. That powerful selective forces are involved is indicated by the fact that their characteristics become more accentuated on approaching the line where they meet. It seems that in these areas the species has reached an adjustment to its environment by different genetic paths. Consequently intermediates between the Cornish and the normal English populations, being at a disadvantage, tend to be eliminated, so accentuating their distinctive features. That this butterfly has indeed built up alternative adaptations to similar situations is indicated by a remarkable fact: the line of demarcation between the two types of spotting is not a fixed one but has, in recent years, advanced eastwards across England and may do so by as much as 20 miles in a generation.

The nature of the differences involved is largely unknown. However, one aspect of them has been detected by McWhirter and Scali (of the University of Pisa, who worked at Oxford for a year). The larvae of this butterfly feed upon grasses of various species. These are, of course, infected with a wide range of bacteria. McWhirter and Scali find that in a majority of the larvae in the Isles of Scilly only Gram-positive forms are present in the mid and hind-gut throughout larval life. Gram-negative types occur in the fore-gut for 20 minutes or so after feeding but are eliminated during that time. The situation is similar in the Winchester district up to and including the fourth larval instar; on the other hand, the mid-gut of the larvae there may either contain only Gram-negatives during the fifth instar or else a mixture of Gram-negatives and Gram-positives during this, the last, larval stage.

It seems then that in different regions the larvae have adjusted themselves to bacteria in distinct ways by producing diverse anti-bacterial substances, apparently in the secretion of the salivary glands, though their nature is still under investigation (McWhirter and Scali 1966).

Genetic diversity is indeed so great that organisms tend to solve the same problem by different means when confronted with it independently in distinct areas, and of this *Maniola* evidently provides examples. As pointed out by McWhirter, this generalization may well explain an important situation sometimes encountered in cultivated plants and domestic animals, for these may produce inferior types when crossed with individuals adjusted to conditions elsewhere which seem identical with their own.

Among the various conclusions to be drawn from the ecological genetics of *Maniola jurtina*, one is outstanding: the exceedingly powerful selection-pressures that operate upon its natural populations. Where the data are available, the strength of the selection-pressure can be calculated, but its existence is to be deduced also from two situations which emerge whenever appropriate studies on this or any other organism are undertaken: the rapidity with which each species adjusts to new conditions and the extreme localization of such adjustment when the environment suited to it is itself sharply circumscribed. Further examples of this may be cited from work upon plant material undertaken by Bradshaw and his colleagues.

They have studied the evolution of tolerance to heavy metals that can be detected in plants growing upon soil containing such substances. There is at least some evidence to show that these adaptations are rapid and certainly they are strictly limited to the contaminated soil.

Snaydon showed in 1962 that the grasses *Festuca ovina* and *Agrostis canina* growing at the foot of galvanized wire netting in the Breck country were significantly more zinc-tolerant than were the same species 6 feet away. This fence was erected in 1936 and renewed in 1958 (see Bradshaw, McNeilly, and Gregory 1965, p. 334).

Bradshaw and his colleagues (see Bradshaw, McNeilly, and Gregory 1965) have examined the extent to which plants growing on disused mine tips are able to withstand the high concentrations of metals sometimes found in the ground at such places. The grass *Agrostis tenuis* proved to be highly copper-tolerant throughout a contaminated area of 2 × 1 miles in Anglesey and to an equal degree on the spoil tip of the Drws-y-coed mine, only 300 × 100 yards in extent. The plants were normal with respect to this quality 7 yards outside the Anglesey area and nearly so one yard beyond that at Drws-y-coed, where tolerance was completely lost 15 yards away. Indeed, it has been shown that the tolerant plants are at a slight disadvantage compared with normal ones when growing on ordinary soils. Some species, however, such as the grass *Dactylis glomerata*, which are common in the immediate neighbourhood of these sites, are absent from them. Evidently, they have not been able to evolve the capacity to withstand the copper present in the soil there.

Gregory and Bradshaw (1965) have shown that tolerance is achieved separately for each of a number of heavy metals, copper, zinc, lead; the ability to withstand one substance does not confer ability to withstand others (except in respect of copper and nickel). Yet multiple resistances may be built up in species from sites where a double contamination exists. It is noteworthy that tolerant and non-tolerant plants take up the metals to an equal degree: the compounds containing them must therefore be rendered inocuous within the plant tissues.

Since, therefore, deleterious doses of heavy metals may be rendered harmless by plants in certain industrial or artificial conditions, one might expect to find species adjusted to growing where high concentrations of toxic substances are normally present. That situation is, of course, well known. Thus *Viola calaminaria* is actually restricted to areas where a large amount of zinc occurs. Furthermore, populations of species can adapt to exceptional natural environments; such as *Gilia capitata* when growing in soil on serpentine rocks, which have a high magnesium content (Gregory and Bradshaw 1965).

The power of selection to produce sharply demarked micro-evolution is thus as clearly indicated by these plant examples as by *Maniola jurtina.* In the butterfly, however, the evidence for the rapidity of the process is far more complete. The development of zinc-tolerance under the wire fence in Breckland could have taken a quarter of a century, though it may well have been much more rapid. It is not known at what rate tolerance to different heavy metals present in mine tips has developed. These spoil heaps have generally been in existence for a century or more. Indeed, as Bradshaw and Gregory indicate, there is not much to be gained by ascertaining their exact age, for the soil over these areas may have contained an excess of the metals in question before the mining operations started.

However, there is clear evidence for rapid evolution under extremely powerful selection in some other forms. Doubtless, the instances are few merely because the researches necessary to demonstrate this situation have seldom been carried out. Certainly the rapid development of tolerance for insecticides provides a further example of them: two years in house flies and mosquitos subjected to D.D.T. (Brown 1957). This has caused great difficulties to those engaged in pest control. The general attitude seems to be that it is necessary to produce a succession of suitable destructive agents so that a new one is always available at any locality as tolerance develops to the compound employed there. It would in reality be more efficacious to use two non-interacting insecticides simultaneously. The fact that plants are found to build up tolerance to two toxic substances present together in the soil is not in itself an argument against that technique, for we have no indication how long such multiple adjustment has taken to evolve. On general grounds, it is likely to be considerably slower than successive adaptations.

Polymorphism

Polymorphism is a distinct and very frequent form of variation, which has been precisely defined in the following terms (Ford 1940): 'The occurrence together in the same habitat of two or more discontinuous forms of a species in such proportions that the rarest of them cannot be maintained merely by recurrent mutation'.

That definition becomes easily intelligible and informative when analysed. We are faced here with discontinuous variation involving clear-cut forms, or 'phases', as with the blood groups or with sex; but evidently the human population is not polymorphic for variations in height.

The phases occur together in the same habitat. That is to say, we are not here concerned with seasonal changes or with geographical races. It will be realized also that polymorphism excludes the segregation of rare major genes to produce disadvantageous abnormalities eliminated by selection and balanced only by mutation. Thus the occurrence of albinos does not make the human race polymorphic for albinism.

Discontinuous variation must clearly be controlled by some 'switch-mechanism', responsible for the development of each individual as one or another type. This control is nearly always supplied genetically. Environmental differences are hardly ever responsible for it, partly because it is difficult for them to produce clear-cut distinctions. It is true that in certain instances dissimilar seasonal forms are known and these must be environmental, as in the spring and summer broods of some butterflies; but here the whole generation is similar. Far greater difficulties would be involved in evoking several distinct types within one brood by environmental means. Still more would it be difficult to ensure that their proportions should be accurately adjusted to meet changes in the habitat or in the hereditary composition of the organism, as is essential in polymorphism.

When a gene mutates, three possibilities confront us; for the new allele to which it gives rise may be either harmful, beneficial, or of neutral survival value compared with its previous and normal counterpart. If harmful, it will be eliminated by selection and cannot evoke a polymorphism. This will be its most usual condition and fate, because mutations are nearly always disadvantageous. That is obvious, since they occur at random relative to the needs of the individual and one cannot make random changes in a highly organized system and expect to promote its harmonious working save on very rare occasions.

Secondly, many polymorphic phases are identified by quite trivial features which, it might be supposed, are of no importance to the organism. Yet it would be false to conclude that the mutants that they give rise to them are of no importance also. On the one hand, genes have multiple effects. We identify them by some feature easily recognized, however trivial, but their more important consequences are less obvious; differences in viability or fertility, for instance, are almost always associated with them. Moreover, Fisher (1930*a*) has shown that the balance of advantage and disadvantage between a gene and its allele must be remarkably exact if it is to be effectively neutral. He has demonstrated also (1930*b*) that the spread of a 'neutral' gene, when such does arise, is quite extraordinarily slow: so slow, indeed, that if derived from a single act of mutation the number of *individuals* carrying it cannot be much greater than the number of *generations* since its occurrence.

The result of these considerations is this: that if a feature controlled by a major gene, apart from polygenic characters that is to say, occupies any appreciable proportion of a population, 1 per cent for example, then that gene, however insignificant its manifestation, must be of importance to the organism. The impact of that conclusion upon medical work and human affairs in general is indeed considerable.

Because a gene is advantageous it will not on that account alone evoke a polymorphism. It will, if its advantage persists, spread through the population until its former normal allele is converted into a rare mutant. During that process a 'transient polymorphism' is produced which, in the end, becomes extinguished. A 'permanent polymorphism' arises only when a condition possesses some advantage when rare; one that wanes and is converted into a disadvantage as it becomes commoner, until the gene controlling it is poised in the population by opposed selective forces. It may be that we are dealing with a feature in which diversity itself is desirable, as with sex, in which an undue excess of males or of females would be opposed by selection. Alternatively, and far more usually, it may be that the heterozygote gains an advantage compared with either homozygote and so maintains diversity. There are two methods by which such heterozygous advantage is achieved. They involve a number of technicalities that need not detain us here. Those who wish to understand them will find them described and explained in *Genetic Polymorphism* (Ford, 1965*b*, pp. 26–8).

It should be noticed that if a gene having considerable effects becomes an asset and increases in frequency, its most usual fate is to acquire such heterozygous advantage rather than to spread until it displaces its allele; thus polymorphism is a very common phenomenon. Evidently it involves evolution, in the sense that a feature previously rare establishes itself in an appreciable proportion in an animal or plant population. But it may involve evolution in another sense also, one which many people would think the more impressive: that is to say, the feature itself is subject to change and adaptation. For selection favours those genetic constitutions that bring out the effects of a gene in the most desirable way, enhancing those that are good and minimizing those that are bad. Evolution in these senses has been analysed by the techniques of ecological genetics and it may usefully be illustrated by examples.

The Scarlet Tiger Moth, *Panaxia dominula*, is abundant in isolated colonies in southern England, generally in marshes and along river banks. The species has only one generation a year and flies during the day; its scarlet hind wings marked with black, and bronze-green forewings with cream spots add a tropical touch to the scene. In an area of about 15 acres at Cothill, Berkshire, it has become polymorphic for a gene increasing the black on the hind wings and reducing one of the cream-coloured spots on the front pair, the form *medionigra*.

The nature of this polymorphism has now been analysed. It is found that *medionigra* is at a disadvantage in two respects: its male fertility is reduced, while its survival from egg to perfect insect is only 75 per cent that of the normal *dominula*. These disabilities are balanced by an asset that accrues from differential mating (Sheppard 1952), in which females of the ordinary form prefer to pair with *medionigra* males rather than with those they resemble; a reaction probably due to a difference in sexual scent.

It seems probable that in the exceptional conditions of Cothill, which is a marsh of unusual character, *medionigra* survives somewhat better than elsewhere. Moreover, the mutation-rate of this form may be so low that it only appeared in this locality quite recently. For Cothill is the only place where this variety is known to have established itself by natural means. Moreover, it is still being adjusted there, since it has become significantly more pronounced in appearance during the last twenty-eight years. Professor P. M. Sheppard has also introduced *medionigra* at a known frequency into another isolated colony where it was absent, and there its distinctive features are diminishing. These two trends have previously been induced in the laboratory by breeding for four generations from the more and from the less extreme examples respectively: the first time, perhaps, in which an evolutionary change occurring in nature had been forestalled experimentally.

At Cothill, the numbers of *medionigra* have been recorded from 1939 onwards and this form has been relatively commoner in some years than in others, while the size of the flying population has also been estimated each season. By combining these two pieces of information, it has been possible to calculate that the extent to which this variety fluctuates in frequency is too great to be due to chance. It must be the result of selection varying annually in intensity and direction (Fisher and Ford 1947).

We may take a second example of a somewhat similar kind. Black forms of eighty or more species of moths have become established in certain industrial areas of Britain and there they have largely superseded the normal pale specimens. The matter has been studied with great success by Kettlewell (1961) using, in particular, the Peppered Moth, *Biston betularia*. This is normally a whitish insect marked with black lines and dots so as to resemble a patch of lichen when resting on tree trunks, as is its habit. Its black variety, which is controlled by a single gene, first appeared in Manchester in 1848. By 1890 about 98 per cent of the moths there were black, in part because they are the better concealed from bird predation when resting upon the smoke-begrimed trees of a manufacturing district. Haldane (1956) calculated that this result is consistent with the fitness of the pale specimens being about half that of the heterozygotes, which also are about 10 per cent hardier than the homozygous blacks. It will be noticed therefore that here, as expected, we encounter heterozygous advantage that had evolved in addition to the new cryptic colouring. Today the two kinds of blacks, heterozygotes and homozygotes, are identical in appearance; blackness is 'dominant', as we say. Yet 100 years ago the heterozygotes at any rate were marked with a scattering of white scales and traces of white lines absent today. The black form therefore has evolved, and in the direction of becoming more extreme.

This has been proved experimentally by Kettlewell, who has shown that the original gene for melanism has been replaced by another allele producing more intense blackening. In addition, his crosses with a closely allied American species, *Biston cognataria*, demonstrated that this effect is the outcome of selection acting gradually on the genetic constitution of the insect; for the interspecific hybrids showed a complete breakdown in the working of the gene responsible for the black colouration, so producing intermediates of all degrees.

It will be noticed that in the last example, although a form was spreading because it had adapted its colouring to a new situation, one arising from the effects of industry, heterozygous advantage had also evolved in it.

The influence of this condition may be further illustrated from ecological genetics carried out on snails. A common British species *Cepaea nemoralis* is polymorphic for several characteristics. The shell may be either yellow (greenish when the animal is within) or dark (pinkish or brown). The yellow type is to our eyes the better concealed upon grass and the dark form upon leaf-litter in a wood. The shells may also be bandless or they may bear up to five dark bands. The plain type is the less easily seen on a uniform background and the banded upon a diversified one, such as along a mixed hedgerow.

Our own judgement in these matters is endorsed by an important predator, the Song Thrush. This carries all but the small, immature, specimens to convenient stones in order to break them open. There the remains of the shells accumulate, so that it is possible to ascertain whether the birds are removing a random sample of the snails. This they do not do, for they destroy a greater proportion of those that are less well concealed.

In each locality, then, the inappropriately coloured and marked forms are constantly subject to elimination (Cain and Sheppard 1950). This is reflected in the extent to which their proportions vary with background: under 30 per cent of the shells are yellow in beech woods while up to 85 per cent are yellow on mixed herbage. Yet, owing to heterozygous advantage, the populations do not become uniform in these places. That is to say, the polymorphism is *maintained* on a physiological basis but the frequency of the phases is *adjusted* by differential predation.

The effect of the predators has, moreover, been quantified. Cain and Sheppard (1954) calculated that one snail population near Oxford consisted of approximately 10 000 specimens; and during a period of only 17 days, 863 of the snails (8·6 per cent of the total) in it were destroyed by thrushes. It may be mentioned, however, that in certain rather exceptional types of country, such as chalk downs, the physiological selection outweighs the effects of predation (Cain and Currey 1963).

Many polymorphisms have no external effects. The human blood groups, of so much importance in medicine and anthropology, are of this kind. They fall into a number of independent systems, thirteen of which are now known. Each comprises a set of alternative forms, the presence of which is controlled by major genes.

The polymorphic nature of the blood groups was recognized only in 1942 (Ford). Yet it should now be evident from what has already been said, for they involve discontinuous phases occurring together in the same population; while even the rarer of them are far too common to be of neutral survival value. That fact had, indeed, been appreciated two years earlier (Ford 1940), when it was recognized that the blood groups must have unknown effects for which they are selected. Some of the conclusions to which this view led were published five years later (Ford 1945), for it was then pointed out that the human polymorphisms must often be associated with liability to develop specific diseases.

The truth of that prediction became apparent in 1953. In that year, Aird, Bentall, and Fraser Roberts showed that a significantly higher proportion of those who develop cancer of the stomach belong to group A (of the OAB series). Other such associations have, as expected, followed. It is to be noticed that some of them cannot themselves be of selective importance to Man, but they indicate additional effects of the genes concerned and of a type which might well influence survival. Thus cancer of the stomach is, like the majority of cancers, normally a disease of later life, the onset of which occurs after the age of reproduction. It is possible, however, that this was not so at an earlier stage in evolution (Ford 1949). Other diseases influenced by the blood-group genes may, however, have more direct bearing upon selection; such as rheumatic fever, which is the outcome of a throat infection by type A haemolytic streptococci, one that is less liable to arise in those who belong to blood group O than to the other phases of the OAB series.

Similar conclusions are applicable to further concealed human polymorphisms, such as ability to taste the sulphur compound phenyl-thiourea. This is intensely bitter to those who can detect it but tasteless to about 30 per cent of western Europeans. Its occurrence influences the types of thyroid disease to which both 'tasters' and 'non-tasters' of this substance may be subject.

There are other instances in which the nature of the selective balance maintaining human polymorphisms is more obvious. Thus sickle-cell anaemia is generally fatal to those homozygous for that condition.† Yet the gene controlling it occurs in as many as 40 per cent of the population in certain African tribes owing to the considerable immunity to subtertian

† See Chapter 6, p. 161.

malaria it confers on the heterozygotes who, moreover, are themselves quite healthy. We have here a situation of contending advantages and disadvantages, one that evidently will not be maintained in regions from which malaria is absent.

It has long been recognized that the frequencies of the blood-group types are of importance in anthropology, though unfortunately the equally great value of the other human polymorphisms in this connection has not been generally realized. The fact that they all provide criteria of relationship, because maintained at constant values in each race, has been widely recognized but, strangely enough, for the wrong reason. Indeed in writing on this matter in 1957, I said:

> 'By a curious inversion of logical thought, it was held that their occurrence in distinct and characteristic proportions in the different races of mankind was especially important because the variation involved was selectively neutral. Precisely the contrary is true. The fact that the genes concerned are balanced by selection at optimum frequencies, which differ from race to race, is the one which gives them significance as a criterion of relationship. It does so because in these circumstances their proportions are influenced by the average genotype of the population in which they occur.'

The existence of polymorphism clearly indicates a situation that is of selective and evolutionary importance; as already indicated, much of it is likely to be concealed, in the sense of having no obvious visible effects. Quite recently, however, a great range of such cryptic variation has begun to be analysed both in vertebrates and invertebrates by means of electrophoresis. Protein variants can be made to migrate at different rates in a medium through which an electric current is passing. In this way the variation can be detected, and much of it proves to be polymorphic. Thus an important source of analysable variability has become available in the last few years. Its study will open up a great field of enquiry of both medical and evolutionary interest. Moreover, such concealed polymorphism can be related to other types of diversity with which it may be associated, as in the distinct stabilizations of the Meadow Brown butterfly in different areas, previously mentioned.

Isolated communities

Attention has already been drawn to some of the special features that distinguish isolated communities. A number of others having both evolutionary and economic interest have been studied using the methods of ecological genetics. Two examples of them can conveniently be discussed here. They are chosen because they both involve promising lines of research that have so far received insufficient attention.

In the first place, isolated animal communities tend to be exposed to special dangers if they occupy small areas. Indeed we may expect individuals in such situations to have a higher average mortality than they would in normal conditions. For this there are several reasons.

Evidently the food supply may be inadequate in such places. There are, however, a great many occasions in which this rather obvious danger does not operate; thus insect species are rarely limited by shortage of food. However, two other components acting unfavourably upon small isolated populations of animals may be mentioned. In the first place, danger is inherent in the fact that as the size of a locality is reduced, so its perimeter, on the average, increases relative to its area. Thus a larger proportion of individuals are liable to wander or be blown away from a small than from a considerable tract of land; a hazard to which flying insects are particularly liable. In addition, a large piece of country is more likely than a small one to contain regions where shelter may be found in all conditions of wind and weather.

Such conditions have very seldom been tested quantitatively, even though the means for doing so have been available for many years. In this connection, we may turn our attention briefly to the method by which the numbers of an animal population, whether of snails, insects, or other forms, can be calculated. It is based upon the method of marking, release, and recapture. Suppose 100 specimens are marked, released, and allowed to mix thoroughly with those in the wild. If another sample of, say, 80 is then withdrawn before there is likely to have been appreciable mortality and 10 of this second sample prove to be marked, we calculate the total population-size as $100 \times 80/10 = 800$.

This is a great oversimplification of the situation, since in many forms, especially insects, the population-size in a given locality may change rapidly. If, for instance, we are estimating the numbers of butterflies or moths in their imago, or winged, state, this may rise from nil to a large maximum and decline to nil again in a few weeks. In such circumstances, we generally need to estimate the number of individuals alive per day, also that comprising the total population of the brood over the whole period of its emergence.

For this purpose we require to know not merely whether an insect has been caught before but precisely on what day. This can be done by marking the wings with a waterproof (cellulose) paint, using minute dots that can be varied as to colour and position to distinguish different dates. The method of analysing the results so obtained involves a certain amount of not very complicated mathematics, easily handled with the help of an electric calculating machine (Fisher and Ford 1947, Ford 1953). A step that has to be taken in that process is to estimate the average expectation of life of the individuals, and this may in itself provide information of much value; it is clearly relevant to relating mortality to size of habitat. That comparison was first attempted by Dowdeswell, Fisher, and Ford (1949) in our studies of the Meadow Brown butterfly which, up to 1953, formed three isolated colonies on the island of Tean, Isles of Scilly (p. 179). The two sexes were treated separately owing to their markedly different behaviour.

The three isolated areas that the species occupied differ considerably in size. The necessary comparison between the insects inhabiting them could be obtained by calculating the mortality in each place and dividing this by a constant estimated value. The result may then, for convenience, be multiplied by 100. Choosing a death-rate of 11 per cent daily, which had been derived as a reasonable figure from the data, the survival of the two sexes is shown in Table 8.1. A figure exceeding 100 indicates an observed survival-rate greater than the expected 89 per cent daily, while a figure below 100 shows that survival does not reach this value.

TABLE 8.1

Area of locality (acres)	Relative death-rate	
	Males	Females
5·3	107·3	108·2
20·8	104·7	—
1·8	89·6	74·3

It will be seen that in the males the result was very similar in the two large areas, while the death-rate was substantially greater in the smallest one. It is also evident that the females, treated independently, corroborrated these findings quite well; unfortunately their numbers could not be assessed in the largest site because insufficient recaptures were obtained there.

In short, the death-rate of *Maniola jurtina* is consistent in the two large areas, while the higher value it attains in the smallest one is striking and evident in both sexes. It would not be wise to put too much confidence in a single set of observations such as these, but we have obtained some confirmation of them in another colony of this butterfly. We undertook these studies largely to demonstrate that this technique could provide useful information. It has indeed stimulated interest but not, unfortunately, research, and this type of enquiry remains undeveloped. It will be noticed that it opens up further lines of investigation: thus it would by these means be possible to compare survival-rates in different environments and of different forms of a species.

A second aspect of isolated communities may briefly be considered here. It is provided by the special adjustments to local conditions often found in species at or near the edge of their range, and by the methods they develop for preventing wide out-crossing there.

Evidently plants and animals are faced with exceptional difficulties as they approach the limit of their tolerance towards the extremity of their distribution. Here they often live in isolation because adapted to some distinctive habitat. This may not be one to which they are normally restricted. But by this means such species may develop accurate adaptations that ensure their survival in relatively unsuitable conditions.

For instance, the Ground Lackey moth, *Malacosoma castrensis*, is extensively distributed in Europe. It has a wide tolerance for different types of countryside, inhabiting heaths and woods as well as coastal areas. In Britain, it is at the extreme north-western limit of its range and it is restricted to isolated localities in the Thames estuary and along the coasts of Essex and southern Suffolk. Here it is found only in salt marshes, and has developed special adaptation to life in them. Thus the eggs have become resistant to salt water. They are laid by the females, which seldom or never fly, upon debris that is washed round the salterns at high tides, and by this means the insect is distributed. The larvae feed upon a wide variety of low-growing plants, which they are likely to find anywhere within the area.

Similarly, the Swallowtail butterfly, *Papilio machaon*, is wideranging over many types of country including mountainous areas in Europe and the Near East, the larvae feeding upon a variety of plants, especially wild and cultivated carrot. In Britain it is restricted to the Norfolk Broads and similar sites near them and, until recently, to a single Fen in Cambridgeshire. Here the larvae eat only Milk Parsley, *Peucedanum palustre*. To ensure localization to the areas to which the species is closely adapted, the butterflies have modified their habits. Though powerful on the wing, they seldom wander from the marshes where they live, a restriction quite contrary to their normal behaviour.

Similar adjustments are to be found in plants at the edge of their range. Thus in Britain and north-western Europe the Box Tree can grow only in strongly alkaline soil, though normally it is subject to no such limitation.

In such circumstances, it is evident that the highly adapted races living generally in isolation should not be contaminated from populations that have evolved differently elsewhere. They should indeed be protected from the transport of alien pollen and seeds if plants and from migration if animals.

It is for this reason that plant species are often polyploids at the edge of their range. That is to say, they have multiplied their sets of chromosomes, a condition that produces sterility when crossed with individuals in which the process has not occurred. Moreover, many groups such as the Primulaceae (including Primroses, Cowslips, and their allies), have developed heterostyly. This is a situation in which the flowers are of two or more distinct types borne respectively upon different plants. Fertilization takes place predominantly between the unlike forms, so encouraging out-breeding. Moreover, the heterostyled form can change to a homostyled one, which is chiefly self-fertilizing. This is a frequent adjustment in heterostyled species near the limit of their distribution. Thus the homostyled *Primula scotica*, which grows in the north of Scotland, Orkney, and Shetland, is an adaptation from the more southerly and heterostyled *Primula farinosa*.

There is, however, a difference in effect between these two methods of securing genetic isolation, for homostyly slows down evolution since it reduces genetic variability. Thus it is a condition to promote stability when a successful adjustment to the environment has actually been achieved. On the other hand, individuals belonging to any one level of ploidy (if it be an even number) experience much less reduction in the variation upon which selection can work, for their outcrossing is but little restricted within their own type.

No mechanisms subserving the same purpose as these ‘mechanical’ devices in plants have yet been found in animals. Polyploidy, which is ostensibly available to both groups, can be excluded. Though not so exceptional as at one time supposed, it is in all animal species, except the most sedentary, very rare compared with plants. This is due to the difficulty of establishing polyploids in their initial stages save in forms capable of vegetative reproduction. For here we have a situation in which is produced an individual, itself fertile, though sterile with the normal type that must at the outset surround it.

It seems likely that the great selection-pressures now known to be operating in nature may be able to eliminate intermediates between forms evolving adaptations to distinct local conditions: as demonstrated, for instance, in *Maniola jurtina*. This would mean that where local races come into competition, smaller selective advantages could be employed in plant communities differing respectively in their number of chromosome sets than in animals.

One of the oustanding results of ecological genetics is indeed the discovery that great selection-pressures are normally operating to maintain and adjust the adaptations of organisms in nature. It needs to be recognized, however, that such forces may often be balanced against one another, so that some features may actually have but a small selective advantage when considered over a considerable period of time, even though the selection operating upon them may be very powerful. Yet in this situation, adaptations can be adjusted rapidly to changing conditions in a way that would be impossible if the different selection-pressures involved were themselves of small magnitude. That is to say, as a result of ecological genetic studies it has been found that micro-evolution can take place far more quickly and effectively than had until recently been appreciated.

Of the many conclusions to which ecological genetics leads, one other may be mentioned here. Polymorphism, necessarily a common phenomenon (p. 183), must always be balanced by selective advantages and disadvantages; its existence therefore advertises a situation of importance to the organism however trivial may be the features by which we recognize it. Attention has been directed to the fact that genes have multiple effects and that the more important of these will generally be cryptic because influencing the physiology of the organism. Thus when we encounter a clear-cut distinction in which, for example, several per cent of men and women are incapable of detecting a particular scent, or when we find that some apparently insignificant quality is polymorphic in a plant or animal population, we can be sure that the condition carries with it important implications. What these are will in each instance be a matter for research, directed especially by the principles of ecological genetics.

References

Aird, I., Bentall, H. H., and Fraser-Roberts, J. A. (1953) *Br. med. J.* (1), 799–801.

Bradshaw, A. D., McNeilly, T. S., and Gregory, R. P. (1965) *5th Symposium of the British Ecological Society*, pp. 327–43.

Brown, A. W. (1958) *Monograph Ser. W.H.O.* **38**, 1–240.

CAIN, A. J. and CURREY, J. D. (1963) *Phil. Trans. R. Soc.* B **246**, 1–81.

—— and SHEPPARD, P. M. (1950) *Heredity* **4**, 275–94.

—— and SHEPPARD, P. M. (1954) *Genetics* **39**, 89–116.

DOWDESWELL, W. H., FISHER, R. A., and FORD, E. B. (1949) *Heredity* **3**, 67–84.

FISHER, R. A. (1925) *Statistical methods for research workers*, 1st edn. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1927) *Trans. R. ent. Soc. Lond.* **75**, 269–78.

—— (1928) *Am. Nat.* **62**, 115–26.

—— (1930*a*) *Proc. R. Soc. Edinb.* **50**, 204–19.

—— (1930*b*) *The genetical theory of natural selection.* Oxford University Press

—— and FORD, E. B. (1947) *Heredity* **1**, 143–74.

FORD, E. B. (1940) Polymorphism and taxonomy. *The new systematics* (edited by Julian Huxley) pp. 493–513. Oxford University Press.

—— (1942) *Genetics for medical students*, 1st edn. Methuen.

—— (1945) *Biol. Rev.* **20**, 73–88.

—— (1949) *Heredity* **3**, 249–52.

—— (1953) *Rep. Australas. Ass. Advmt Sci.* **28**, 143–54.

—— (1957) *Nature, Lond.* **180**, 1315–19.

—— (1965*a*) *Ecological genetics*, 2nd edn. Methuen, London.

—— (1965*b*) *Genetic polymorphism* (All Souls Studies, V), Faber and Faber.

FORD, H. D. and FORD, E. B. (1930) *Trans. R. ent. Soc. Lond.* **78**, 345–51.

GREGORY, R. P. and BRADSHAW, A. D. (1965) *New Phytol.* **64**, 131–43.

HALDANE, J. B. S. (1956) *Proc. R. Soc.* B **145**, 303–6.

KETTLEWELL, H. B. D. (1961) *A. Rev. Ent.* **6**, 245–62.

MCWHIRTER, K. G.(1968) *Heredity* (in the press).

——and SCALI, V. (1966) *Heredity* **21**, 517–21.

SHEPPARD, P. M. (1952) *Heredity* **6**, 239–41.

9 *Hormones and transmitters*

By 1920 the ideas on the means by which information is transmitted from cell to cell in a multi-cellular organism could be summarized like this: cells are connected in two ways, either by nervous connections or by means of the blood stream.

The concept of the nervous system as a connecting link between distant parts of the body had been elaborated in the nineteenth century, and the reflex arc had been established as the functional and anatomical link between distant cells. The knowledge of the reflexes was rounded off by the work of Sherrington at the beginning of the present century.

Another pioneer of the study of reflexes, Pavlov, had already investigated the reflex control of secretion in the digestive tract. It is interesting to remember that this work not only initiated, in Pavlov's own researches, the concept of conditioned reflexes, but that it also stimulated the classical work of Bayliss and Starling in which the term 'hormone' was introduced. In 1902, Bayliss and Starling, although they confirmed Pavlov's observations on the occurrence of pancreatic secretion in response to electrical stimulation of the vagus nerve, discovered that the stimulus that normally initiated pancreatic secretion when food passed from the stomach into the duodenum was not dependent upon intact nervous connections between duodenum and pancreas. The stimulus was shown to be carried by an agent released from the mucous membrane of the duodenum that reached the pancreas by way of the blood circulation. In other words, there existed here a substance that acted as a chemical messenger, and for this kind of messenger Bayliss and Starling introduced the new term 'hormone'. The introduction of this new term and of the hormone concept was the birth of what became a new discipline, *endocrinology*.

It has often been pointed out that the new concept was preceded by many observations that antedated Bayliss and Starling. Chandler Brooks has discussed some of these forerunners; he refers to the doctrine of Theophile de Bordeu, enunciated in 1775, which postulated that glands and other organs produced secretions that reached distant organs by the way of the blood stream. An early and elegant analysis of a hormonal effect was given by Berthold (1849) in a study in which he showed that an implanted testis could prevent the signs of castration in cocks, irrespective of the site of implantation.

Similarly, Vulpian and Brown-Séquard postulated the release of substances into the blood stream. The list of forerunners of the endocrinologists could easily be extended. We need only to record the discovery by von Mering and Minkowski, of pancreatic diabetes, and that by Baumann, of organically bound iodine in the thyroid gland.

Into the eighteen-nineties there falls also the discovery by Oliver and Schäfer of the effects upon the blood pressure of extracts of the adrenal medulla. In the decade that followed, this work led to the isolation of the first hormone in a chemically pure state. This was adrenalin, synthesized in 1905.

Thus, before 1920 it had been established that there were two mechanisms by which information could be transmitted to an 'effector' cell. The first of these was a nervous mechanism, the reflex. The second was the hormonal mechanism, the secretion of a stimulating agent that was widely dispersed by way of the blood stream. These two mechanisms were considered as essentially distinct.

The distinction between a reflex and a hormonal response is important: the reflex arc ensures a strict spatial localization of the response: the stimulus is carried only to those cells that are innervated by the efferent link of the reflex arc. In contrast, the second mechanism, internal secretion, relies for its effectiveness on the blood circulation, which reaches practically all cells in the body. If there is any specificity in the response to a hormone this is not achieved by anatomical complexity, as in the reflex pathway; it is achieved by the ability of the effector cell to select, and respond to, specific chemical stimuli.

The universal exposure of most parts of the body to the same chemical environment made almost any substance circulating in the blood stream a chemical messenger. What produces the variety in the response is the differing sensitivity of the cells to these agents. Bayliss and Starling had no hesitation in including carbonic acid among the hormones; at the other end of the scale there was secretin, an example of a hormone to which only a very restricted group of cells responded.

The idea of chemical specificity will be discussed below. Here it is introduced to point to an important practical outcome of the hormone theory, one that followed immediately from the concept of chemical messengers. Adrenalin may serve to illustrate the point. Once the chemical constitution of the hormone had been elucidated and it had been prepared by synthesis, pharmacologists began to study the chemical specificity of the response, and then the organic chemists prepared substances chemically related to adrenalin. Of these substances, some were endowed with biological properties differing to a varying degree from those of the hormone itself. These studies led to the introduction of many new drugs of therapeutic value. In the past sixty years an enormous number of synthetic compounds have been prepared, and of these many are used in preference to the naturally occurring hormones in therapy. In addition, the observations on differing pharmacological properties of synthetic hormone analogues have been the basis of the refinement of the receptor concept, discussed below.

Since 1920, the ideas on transmission of information that were current after the introduction of the hormone concept have undergone profound changes that amount to an almost complete reversal. This reversal was initiated by new discoveries that led to the introduction of the theory of humoral transmission of nervous impulses. This theory, first formulated by Otto Loewi in 1921, was extended and elaborated in the nineteen-thirties, chiefly by Sir Henry Dale and his school.

According to the theory of humoral transmission, the response of an effector organ to a stimulation by way of its nerve is brought about by a release from the nerve ending of a substance that exerts excitatory (or inhibitory) actions upon the effector cells. Loewi proposed this theory when he discovered the so-called *Vagus* and *Accelerans* substances in the frog's heart. His theory was extended, first to the whole of the autonomic nervous system, and later also to the nerves that carry impulses to the voluntary (or skeletal) muscles. Although the transmitter theory has been securely established only for a very small number of nerve fibres with endings in the central nervous system, many pharmacologists and physiologists today think of transmission of impulses in the central nervous system in terms of the transmitter theory. Such ideas, although not strictly proved, are supported by a number of observations. First, there is the fact that the two substances that have been proved to be transmitters in the peripheral nervous system, acetylcholine and noradrenalin, are present in the brain, and second, that the enzymes taking part in the formation and in the inactivation of these substances are present also.

If the theory of humoral transmission of nervous impulses is today considered as valid for most, if not all, parts of the nervous system, we can see in it a counterpart to the neurone concept. It was Cajal who defined the nerve cell with all its processes as the functional unit in the nervous system. The term 'neurone' was introduced for this unit by Waldeyer. The idea, according to which the nervous system was composed of the neurones as functional units, was not at once generally accepted. In the first decade of the twentieth century there were still many neurologists who believed in continuous conduction in the nervous system. 'Continuity' and 'contiguity' were the watchwords of the two opposing schools of thought.

The theory of humoral transmission of nervous impulses gave the concept of contiguity a new meaning. The neurone is seen to act as an 'all-or-none' unit, but where it ends and where the message has to be handed on from one such unit to another, a functional gap has to be bridged. A new mechanism comes into play here and that mechanism is akin to secretion.

If we consider internal secretion and nervous activity in the light of the new knowledge we can say that the main difference between these two mechanisms is that in the transmission of nervous impulses the gap that has to be bridged by the transmitter substance is minute. Moreover, powerful mechanisms exist that normally prevent or minimize the spread of the transmitter from the site of its release. Signalling by way of hormones ensures that all cells of the body are reached, but the response to a transmitter is restricted both in space and time. Thus, one and the same transmitter can serve in a great variety of locations and at very different kinds of nerve ending. It is understandable, therefore, that in mammalian physiology, only two substances have so far definitely been established as transmitter substances at peripheral nerve endings; these are acetylcholine and noradrenalin. There may be a few more but the evidence is incomplete, and these same two substances are likely candidates for transmitters in the central nervous system; there are probably other central transmitters, such as dopamine, 5-hydroxytryptamine (serotonin) and others, but for these substances the evidence is at present less secure.

The contrast between the small number of established transmitters and the great number of hormones becomes understandable in the light of what has just been said. Most hormones reach practically all cells of the body in the same concentration (although there are a few interesting exceptions to this general rule). It is the indiscriminate kind of exposure that again emphasizes the need for specificity in the response evoked in the effector cell. Specificity in the response to the transmitters is not unimportant: it is interesting to note a number of instances where the two known transmitters act upon the same tissue but with antagonistic effects. Release of one transmitter may result in excitation, that of the other in inhibition.

The narrowing of the gap between hormone and transmitter concepts that we have witnessed in the past forty years is reinforced by a number of findings. First, let us again consider adrenalin. There is a close chemical and developmental relationship between the cells of the adrenal medulla, which contain both adrenalin and the closely related noradrenalin, and the so-called adrenergic neurones, which make use of one or the other of these two substances as transmitters in different species. The cells of the adrenal medulla and the nerve cells of the adrenergic neurones develop and migrate from the same area in the embryonic vertebrate and acquire their differing shapes and functions in subsequent stages of cell differentiation. Also, these cells are similar in their equipment with chemical tools: of all the nerve cells, only the adrenergic ones seem to contain all the enzymes required for the synthesis of the adrenergic mediators. Secondly, there is the concept of *neurosecretion*. Neurosecretion is exemplified by the mode of release of the posterior pituitary hormones. According to present concepts, the posterior pituitary hormones, oxytocin and vasopressin, are made by nerve cells situated in the brain, in the hypothalamic region. The hormones are then believed to travel in the axonal processes of these cells, down the pituitary stalk. The hormones are stored in the nerve endings, which are situated in the posterior lobe of the pituitary gland, and they are released from these endings when the neurones carrying them are stimulated, presumably in the brain. The hormones are stored in cell organelles, granules that can be seen by the microscopists and electron microscopists; in these organelles the hormones are believed to be associated with the so-called 'neurosecretory material'. The terminology used by students of neurosecretion illustrates clearly the extent to which the transmitter theory has influenced contemporary thought: the posterior pituitary hormones are seen as true hormones, as chemical messengers released into the blood-stream, but they are released, like the transmitters, from nerve endings when the nerve is stimulated.

An analogy between hormones and transmitters that is being revealed in present studies follows a discovery by Katz and his colleagues, who have found that the so-called cholinergic transmitter, acetylcholine, is released in packets. The theory of the 'quantal release' of acetylcholine at motor nerve endings has its counterpart not only at adrenergic nerve endings but also in observations on the release of hormones. Each of the cell organelles that stores hormone in the posterior pituitary gland can be seen as representing a 'quantum' of chemical messenger. As in the nerve endings, the release of messenger from a single secretory granule is pictured as an 'all-or-none' event. In recent years much has also been learnt about the release of hormone from the adrenal medulla; here there is good evidence in favour of the view that secretion occurs by the emptying of a storage granule, so that here too a 'quantum' of hormone is set free when the secreting cell is stimulated. Katz has proposed that the counterpart of the secretory granules seen in endocrine cells in the motor nerve endings are the so-called 'synaptic vesicles', structures that have been revealed by electron microscopy in a great variety of nerve endings.

Chemotherapy and antimetabolites

Chemotherapy as we know it today owes its theoretical basis mainly to the work of Ehrlich.

The nineteenth century had seen the rise of microbiology, the scientific study of microorganisms. The foundation of this science was laid by Pasteur, and his work was followed by that of Koch, who devised methods for growing bacteria in culture and discovered the organisms that caused anthrax, cholera, and tuberculosis. Ehrlich's early contribution had dealt with the means of making these microorganisms visible under the microscope. This was part of his work on microscopic staining techniques, which led him to postulate that chemical compounds, in order to interact with cell constituents, had to be bound by chemical forces to some structural element or substance present in that cell. When at a much later stage he turned his attention to drug treatment of infectious diseases he postulated again the presence in the cell of 'chemoreceptors' that specifically interacted with the chemotherapeutic agent. I quote from a late review of his work, from the *Proceedings of the* 17*th International Congress of Medicine*, held in London in 1913, where he says: 'If the law is true in chemistry that *corpora non agunt nisi liquida*, then for chemotherapy the principle is true that *corpora non agunt nisi fixata*.' The concept of receptors, which stems from Ehrlich's earlier work, will be discussed below; here we just refer to his conclusion 'that in the parasites there are present different specific chemoreceptors: for example, an arsenoceptor which anchors the arsenic radicle, an acetoreceptor which binds to itself the acetic acid residue . . ., and many others. The complete and exhaustive knowledge of all the different chemoreceptors of a certain parasite I should like to designate as the therapeutic physiology of the parasite cell, and this knowledge is a *sine qua non* for success in chemotherapy.' Ehrlich surmized already that the drugs were bound to receptors designed for normal metabolites. The aim of chemotherapy, the '*therapia sterilisans magna*', was a rapid and selective killing-off of the invading parasites. The drugs to be used were those for which the host cells had no receptors.

Ehrlich's own contribution to applied chemotherapy was one of the first drugs that attempted to fulfil these conditions. This was salvarsan, a drug that proved useful in the treatment of syphilis and a number of related diseases. However, the selectivity of the drug was not complete and its toxicity high.

It was only a quarter of a century later that with the introduction of the sulphonamides and the antibiotics truly selective drugs became a reality. The picture of drug action that we owe to Ehrlich is directly applicable to the mode of action of these drugs. For instance, penicillin interferes with the chemical processes that occur during the laying down of the bacterial cell wall. Since the substances present in bacterial cell walls do not occur in the cells of mammals, the mammalian cells do not contain the enzymic equipment required for the formation of bacterial cell walls. Thus, the receptors for the drug are present in the bacterial cells but absent from the cells of the host.

For present-day theories of the mode of action of chemotherapeutic agents the analysis of the action of sulphanilamide has been even more important. In 1940, Woods isolated from bacterial extracts a substance that counteracted the growth-inhibitory effect of sulphanilamide on living bacteria. This substance was shown to be para-aminobenzoic acid. The idea that chemotherapeutic agents acted by interfering with the utilization of substances essential in bacterial metabolism was at that time being discussed in the laboratory of Sir Paul Fildes, where Woods was working. Thus, the theory was put forward that para-aminobenzoic acid was an 'essential metabolite', as defined by Fildes, of bacteria and that the sulphanilamide exerted its effect by a competition with this normal metabolite.

Like penicillin, sulphanilamide fulfils the conditions for selectivity laid down by Ehrlich: it interferes with the enzymes essential for the incorporation of para-aminobenzoic acid into pteroylglutamic acid. In man, pteroylglutamic acid is a vitamin; it cannot be built up from smaller building elements and the enzymic equipment sensitive to sulphanilamide is absent in man. Here again, we meet the phenomenon of specificity: the molecule of the drug is effective because it resembles, in shape and other qualities, the essential metabolite with whose utilization it interferes.

Receptors

The concept of receptors has its origin in Ehrlich's immunological studies. His side-chain theory determined his ideas on chemotherapeutic agents and it also had a strong influence on the theoretical basis of contemporary pharmacology.

Ehrlich's first contributions to immunology were concerned with toxin and antitoxins. In his side-chain theory he postulated that a bacterial toxin attached itself to the cells upon which it acted. This fixation took place by an interaction between toxin and a specific receptor on or in the cell, a 'haptophore' group. In cells sensitized to the toxin by immunization, these haptophore groups were produced in excess; they were secreted from the cells into the blood plasma where they circulated as antibodies (antitoxins). Toxin and antitoxin interacted by virtue of their fit. This picture was fully developed in 1900, when Ehrlich delivered the Croonian Lecture to the Royal Society. In this lecture he referred to the picture of the lock and key, used for the enzyme-substrate interaction by Fischer a few years earlier (see below).

In pharmacology the concept of receptors for hormones and what we would today call transmitter substances, as well as for the so-called 'blocking agents' for these compounds, was first used by Langley. He assumed that the cells responding to drugs like pilocarpine or adrenaline contained what he called a *receptive substance*. In his 1905 paper he wrote: 'I conclude then that in all cells two constituents at least must be distinguished, (1) substance concerned with carrying out the chief function of the cells, such as contraction, secretion, the formation of special metabolic products, and (2) receptive substances especially liable to change and capable of setting the chief substance in action. Further, that nicotine, curare, atropine, pilocarpine, strychnine and other alkaloids, as well as the effective material of internal secretions produce their effects by combining with the receptive substance and not by a direct action on axon-endings if these are present, nor by a direct action on the chief substance.'

Langley was aware of Ehrlich's work on immunoreceptors. He wrote: ' . . . on the general lines of Ehrlich's immunity theory, it might be supposed that a receptive substance is a side chain molecule of the molecule of contractile substance, but at present there does not seem to me to be any advantage in attempting to refer the phenomena to molecular rearrangement.'

Langley's work, in its turn, did not go unnoticed by Ehrlich. It influenced him in his later studies on chemoreceptors. In his London lecture of 1913, already quoted, he writes: 'Earlier studies conducted in a different field, that of the toxins and antitoxins, pointed to the nature of these processes. It was found that the toxins exert their injurious action on the cell by virtue of the fact that they are taken up by certain specific components—side-chains—of the cell, which I have called receptors, and that the antibodies represent nothing more than the cell receptors produced in excess under the influence of the toxin and thrust off.'

'For many reasons I had hesitated to apply these ideas about receptors to chemical substances in general, and in this connection it was, in particular, the brilliant researches of Langley, on the effect of alkaloids, which caused my doubts to disappear and made the existence of chemoreceptors seem probable to me.'

Ehrlich's initial hesitation concerning the use of the term *chemoreceptor* (I find it used for the first time in his second Harben Lecture for 1907) is described more fully in his earlier papers. He did not see the same active response, as in immunization, when he considered the reaction of the tissue cells to the alkaloids; in particular, there was no antibody formation. In a lecture given to the German Chemical Society in 1909 he quoted Hans Meyer, who had pointed to the reversible character of the interaction between alkaloids and tissue 'receptors'; the recognition of the reversibility of this interaction seems to have played a part in his change of mind.

Langley's ideas on the receptive substance made their entry into pharmacology more slowly than Ehrlich's ideas on the immunoreceptors and the chemoreceptors. They were taken up by Clark (who had worked in Langley's department in Cambridge). He formulated his position more concisely in 1937, in a book entitled *General pharmacology*, where we read as follows.

'The author feels he owes an apology for introducing somewhat vague speculations. It is however very difficult for the human mind to function without some form of working hypothesis. The adoption of the conceptions outlined regarding the structure of the cell surface does help to suggest a means by which minute quantities of drugs can produce a highly selective action. Furthermore the author feels that it is very difficult to explain such actions unless one adopts the hypothesis that cellular activities are dependent on receptor groups arranged in some pattern on the cell surface and that the drugs produce their effect by combining with these receptors. This hypothesis is, of course, very similar to that put forward by Ehrlich a quarter of a century ago.'

Since the publication of Clark's book the receptor concept has established itself in the pharmacological literature. There still has, however, been occasional hesitation to use the term. In 1943, Dale, writing in an introductory address to a Faraday Society discussion on 'Modes of drug action', said: 'It is a mere statement of fact to say that the action of adrenaline picks out certain effector-cells and leaves others unaffected; it is a simple deduction that the affected cells have a special affinity of some kind for adrenaline; but I doubt whether the attribution to such cells of "adrenaline-receptors" does more than re-state this deduction in another form.' It is interesting that in the intervening years the analysis of the action of adrenaline in particular has done most to naturalize the receptor idea in pharmacology.

For the modern attitude, let us quote from an introductory article by de Jongh to *Molecular pharmacology* by E. J. Ariëns (1964): 'To most of the modern pharmacologists the receptor is like a beautiful but remote lady. He has written her many a letter and quite often she has answered the letters. From these answers the pharmacologist has built himself an image of this fair lady. He cannot, however, truly claim ever to have seen her, although one day he may do so.'

Chemical specificity

The interactions that have been discussed here, that between a chemical messenger, be it a hormone or a transmitter, or that between an antigen and an antibody, are examples of the phenomenon of chemical specificity. This is a concept that has been found to apply in many branches of biology. It is inherent in the Crick–Watson theory of deoxyribonucleic acid (DNA) replication, the theory that has given us a picture of gene replication.† The same principle is operative in all stages that lead from DNA to protein synthesis. It is through the ideas that have been first put forward by the molecular biologists that chemical specificity is implied in current theories of memory.

Chemical specificity of biological processes was first postulated, in the nineteenth century, in the study of enzymic processes.‡ In the course of that century it became gradually clear that enzymes were distinctive entities and that they could be dissociated from living cells. Although Pasteur had opposed this idea until very late in his lifetime, his own early research had in fact been very important in the development of the idea of specificity. In his early work, Pasteur had shown that micro-organisms were able to discriminate between the two stereoisomeric forms of an organic substance. These early observations were extended towards the end of the century by Fischer and Thierfelder (1894): they showed that yeast cells possessed chemical specificity. These cells were able to distinguish between different groups of sugars, each group composed of compound that belonged to one stereochemical series. They suspected that this power of the cells to discriminate was due to the fact that in the cell the sugar interacted with a protein. Since proteins were also optically active substances, that is, substances that contained centres of asymmetry, they assumed that those sugars that could be fermented were those able to interact with a protein that in its geometry did not differ too much from that of glucose, the prototype of a sugar readily fermented.

In a subsequent paper, written in the same year, 1894, Fischer showed that group specificity was not only a property of intact yeast cells; it was a property also of the enzymes isolated from them. He concluded that the selective action of two enzymes, emulsin and invertin, on glucosides could be explained if it was assumed that only for a similar geometrical configuration could the molecules of enzyme and substrate approach each other so that the chemical process could be elicited. This argument is immediately followed by the often-quoted passage that runs in German as follows: 'dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen, um eine chemische Wirkung auf einander ausüben zu können' (translated, 'that enzyme and glucoside must fit on to each other like lock and key in order to exert a chemical effect upon each other'). This passage is immediately followed by the statement that thus a phenomenon has been removed from the realm of biology and transferred to that of chemistry.

† See Chapter 7, p. 161f.

‡ For the study of enzymes from a chemical point of view, see Chapter 6.

Fischer's guess, that the enzymes are proteins and that the phenomenon of substrate specificity is bound up with protein structure, has been amply confirmed by later work. Today it is strange to recall that as late as 1926 Willstätter, the great chemist who had elucidated the structure of many naturally occurring substances, after many years spent on investigating the properties of enzymes (work that resulted in the introduction of many methods of enzyme purification still in use today), concluded that enzymes were probably not proteins but low-molecular-weight compounds attached to macromolecules. It is one of the ironies of scientific history that from the same year 1926 dates the first successful purification and crystallization of an enzyme, urease, by Sumner. This and subsequent work established the protein nature of the enzymes.

Fischer's discussion of specificity was restricted to the problem of fit; the spatial relationship between substrate and enzyme was emphasized. The question as to the forces at work in the interaction of the two reactants remained open.

The picture of the lock-and-key relationship introduced by Fischer was further discussed by Haldane in 1930. He accepted the picture but discussed also its limitations. He wrote: 'Using Fischer's lock and key simile, the key does not fit the lock quite perfectly but exercises a certain strain upon it.'

Still, the picture was useful. In 1928, Quastel and Wooldridge discovered the inhibition of the enzyme succinic dehydrogenase by malonic acid. Malonic acid was considered to attach itself to the active centre of succinic dehydrogenase by virtue of its structural resemblance to succinic acid. In this way it prevented a molecule of substrate, succinic acid, attaching itself to the enzyme. This inhibitory effect of malonic acid could be overcome by increasing the concentration of succinic acid; this also was an observation made by Quastel and Wooldridge. For this type of inhibition we still use the term 'competitive inhibition', one that appears to have been first used by Haldane.

It seems probable that in the drug–receptor (or hormone–receptor or transmitter–receptor) and antigen–antibody interactions specificity is based on the same principles as in the substrate–enzyme interaction. The early workers left open the question of the forces active. It was from the immunologists that the first studies of the chemistry of specificity came. This work was summarized by Landsteiner in his book *The specificity of serological reactions*. It is fitting that the revised edition of this book published in 1945, two years after Landsteiner's death, contains an appendix by Pauling, whom Landsteiner had invited to contribute an essay on 'Molecular structure and intermolecular forces', in which the forces acting between macromolecules were discussed for the first time. The chief interest, in relation to what has been discussed here, is that Pauling stressed the steric factors; one might say that he gave chemical content to the picture first used by Fischer.

Fischer's picture of the lock and the key perhaps stressed the rigidity of the system responsible for the fit. In the field of enzyme chemistry, recent ideas have stressed the plasticity of enzymes. These ideas culminate in the concept of allosteric interaction. In the theory of Monod, Changeux, and Jacob, enzymes that are important in regulating important metabolic steps have, in addition to the specific site that interacts with the substrate, one or more other specific sites, the so-called 'allosteric sites' that interact with the protein and act as stimulators or inhibitors. They cause a change in the shape of the enzyme molecule which affects the affinity of the enzyme for the substrate; in this way the activity of the enzyme is influenced.

It is not surprising that our knowledge of chemical specificity is still largely based on the study of the enzyme–substrate relationship. Like the receptors for hormones and transmitters, the chemistry of the antibody is still imperfectly understood. However, promising beginnings have been made in the study of the chemistry of antibodies, and it is known which part of the protein molecule is the carrier of specificity. This makes one hope that the interaction between antibody and antigen will eventually be defined in chemical terms. The ideas on the chemical reaction between biological macromolecules exhibiting specificity will have to be assessed by a historian of the future.

Selected References

CLARK, A. J. (1959) *Applied pharmacology*. London.
DALE, H. H. (1934) *Br. med. J.* 835–41.
EHRLICH, P. (1913) *Proc.* 17*th int. Congr. Med.*, p. 505.
LANDSTEINER, K. (1946) *The specificity of serological reactions.*
LANGLEY, J. N. (1905). *J. Physiol.* **33,** 374.
LOEWI, O. (1921) *Pfluger's Arch. ges. Physiol.* **189**, 239–42.
STARLING, E. H. (1905) *Proc. R. Soc.* B77, 505,
WALDEYER, W. (1891) *Dt. med. Wschr*, **17**, 123–8.
WOODS, D. D. (1940) *Br. J. exp. Path.* **21**, 74–90.

10 *Cell Biophysics*

THE development of cellular biophysics in the first half of this century has been based largely on the detailed application to biological problems of concepts derived from two of the physical sciences: physical chemistry, particularly electrochemistry and thermodynamics, and electricity, particularly the physics of transmission lines. The biological problems to which these concepts have been applied form an important field of physiology and they include the problems of nerve and muscle excitation and muscle contraction—problems that have excited the common interest of both physical and biological scientists for a long time since the simultaneous discovery of electricity and 'animal electricity' arising from the famous controversy between Galvani and Volta. It is impossible within the space of one chapter to discuss the developments in the whole of this field and I shall therefore restrict the discussion to one particular topic, that of nerve excitation, in the belief that to give one detailed example will give a more useful impression of the development, than to give a more superficial account of the whole field.

Electric current flow in excitable cells

Before discussing the developments in the first half of the twentieth century, it will be necessary to review briefly the state of knowledge in the relevant fields of physical science at the turn of the century. The theory of transmission lines had been developed by Lord Kelvin in 1856. The transmission line is basically a length of conductor (characterized by its resistance per unit length, *R*) separated from a conducting medium (in the case of a submarine cable, for example, this medium would be the sea) by an insulating sheath with various electrical properties. Voltage signals are applied to one end of the line and are transmitted to the other end by current flow along the core conductor. This current, I, is generated by a voltage gradient according to Ohm's law:

$$\partial V/\partial x = -IR, \tag{1}$$

where V is the voltage of the signal at each point, x, along the line. As it flows along the line, the current is distorted by leakage through the insulating sheath, the density of the leakage current, I, being given by the rate of change of I along the line (Kirchhoff's law):

$$\partial I/\partial x = -I_1. \tag{2}$$

If the sheath can pass a steady leak of current (determined by the sheath conductance, G) and can also distort the signal by drawing transient currents from it in the form of charge stored on the sheath capacity, C, then the leakage current will be given by the sum of the capacitive and resistive currents flowing through the sheath (cf. Fig. 10.1):

$$I_1 = C\,\partial V/\partial t + GV, \tag{3}$$

where t is time. The basic differential equation for this transmission line is obtained by combining equs (1)–(3) to give

$$\frac{1}{GR}\,\partial^2 V/\partial x^2 - \frac{C}{G}\,\partial V/\partial t - V = 0. \tag{4}$$

If C, R, and G are constants (as they usually are in a simple metallic cable) the system is said to be linear and equ (4) may then be solved, for various kinds of signal applied to the transmission line, by conventional mathematical techniques. Some of these techniques have in fact been developed relatively recently. However, the general behaviour of a system obeying equ (4) was well enough known by the turn of the century for the transmission line to serve as a useful model.

The resemblance of nerve and muscle cells to transmission lines was noted at this time by several workers, and Cremer and Hermann drew attention to some of the important physiological consequences of applying the equations to excitable cells. In order to do this it is necessary to identify the elements of the model with excitable cell components. Although no good direct evidence was available until two or three decades later, it seemed reasonable to identify the internal cell fluid as an ohmic conductor carrying the signal currents and the cell surface (or membrane) as a leaky capacitor carrying the leakage current.

Although this analogy between transmission lines and excitable cells has proved extremely useful, it is also important to note its limitations. These arise not so much from the fact that excitable cells are not simple transmission lines (in fact, when they carry only small signals, they behave quite accurately as linear transmission lines) but rather from the large quantitative differences between the electrical characteristics of excitable cells and those of man-made transmission lines. The biological equivalent of the core conductor is a rather dilute salt solution whose conductivity is far smaller than that of conducting metals. Moreover, nerve and muscle cells are extremely thin (commonly of the order of 0·001 mm). The combined result of these two factors is that the resistance of only one or two millimetres of an excitable cell may be equal to that of a whole transatlantic cable. Despite the fact that nature has succeeded in evolving a fairly good insulating sheath (formed by a fatty material, myelin) the tendency for current to leak across it is very much greater than in a man-made cable. The resulting distortion of the signal current is so great that a signal would be reduced to a negligible size within a distance of only half a centimetre if the only method of transmission were that described by linear transmission line theory. Signals are in fact transmitted across very much greater distances in the nervous system, which leads to the conclusion that a different, or greatly modified, transmission process must occur in nerve and muscle cells. One of the major controversies in the field during this century has centred round this question, and some physiologists have favoured the view that the transmission process is less purely physical than processes of the kind described by transmission-line theory, involving perhaps the release from

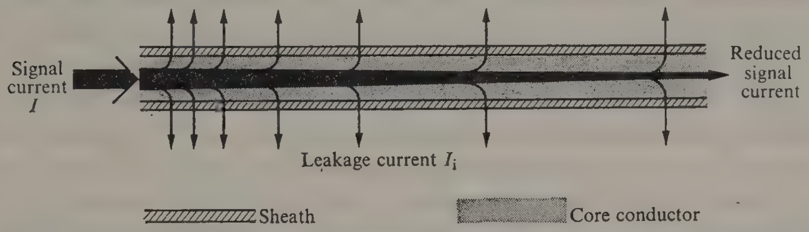


FIG. 10.1. The transmission-line model. The signal current, I, is applied to the core conductor at one end and is conducted to the other end. Leakage of current through the sheath reduces the signal progressively. In a transatlantic cable it would require many hundreds of miles to reduce the signal to the extent shown. In nerve and muscle cells this degree of decrement would occur in about half a centimetre.

or within the cell or cell membrane of chemicals that have the effect of exciting inactive parts of the cell to produce more chemical. In this way a chemical wave might be transmitted from one end of a nerve cell to the other. This view became popular at one time (particularly between the two world wars) for two reasons. First, the electrical transmission theorists could not at that time produce a completely satisfactory modification of transmission-line theory that allowed a rigorous demonstration that the chemical wave hypothesis was unnecessary. Hermann had in fact shown that the current generated by a cell may be sufficient for electrical propagation, but the argument was not quantitative enough to exclude the possibility that the process might be helped to a greater or lesser extent by a chemical wave. Second, it became established during this time that the most likely mechanism by which excitation is transmitted *between* excitable cells (e.g. from one nerve to the next, or to a muscle fibre) does involve the release of an excitatory chemical, which became known as the *chemical transmitter*. This theory was established by Loewi working on transmission between nerves and heart muscle, and by Dale and his colleagues working on the transmission between nerves and skeletal muscles.† The idea that a mechanism of the same kind also operates in the transmission of excitation along individual excitable cells was therefore attractive. This kind of argument can, of course, work both ways: some physiologists remained convinced for some time that the transmission between cells is electrical, partly because they were attracted by the idea that the electrical mechanism is the only method of transmission.

The electrical transmission theory was eventually established in the 1930s in two ways. First, Osterhout and Hill demonstrated that transmission could not occur if electric current flow between two areas of an excitable cell is prevented by allowing part of the cell to pass through an insulating air phase. Transmission could be easily re-established if the non-conducting air phase was short-circuited electrically. Second, Hodgkin showed that the electric current flow is adequate for propagation. In order to describe the theoretical and experimental work that led up to these crucial experiments it will, however, be necessary to return to the situation at the turn of the century and to describe some developments in physical chemistry and their application to excitable cells, since the eventual success of the electrical transmission theory resulted largely from the modification of classical transmission-line theory to include the complex electrochemical properties of excitable cells.

† See Chapter 9, p. 198.

Electrochemical properties of excitable cells

The fluids in biological systems were known to be dilute salt solutions (also containing, of course, many organic molecules) and the elementary physical chemistry of such solutions and, in particular, their properties in the presence of selectively permeable membranes of the kind that may be formed by the cell membrane, was only just beginning to become clear at the turn of the century. The establishment of the ionic theory of salt solutions was of crucial importance. According to this theory, a salt such as sodium chloride (NaCl) when dissolved in water separates into two charged species called *ions*. The sodium acquires a positive charge and becomes the cation, Na^+, while the chloride acquires a negative charge and becomes the anion, Cl^-. This theory explained a number of otherwise puzzling properties of salt solutions and, in particular, it explained why they can conduct electric currents relatively easily. An electric current consists in the flow of charged species. In metals, for example, this flow is thought to be the movement of free negatively charged electrons between the atoms of the metal. In salt solutions, the current may be carried by both ionic species moving in opposite directions, a movement of positive charge in one direction being equivalent electrically to a movement of negative charge in the opposite direction. Excitable cells, however, not only conduct electricity; they also generate it. Thus, it was known at the end of the nineteenth century that an electric current would spontaneously flow between damaged and intact areas of an excitable cell and this current became known as the injury current. The possible mechanisms underlying this property became clearer when Planck and Nernst developed thermodynamic and kinetic theories for the origin of electrical potentials in salt solutions. Although all ions carry the same charge, or integral multiples of this charge, they do not move with the same degree of ease through solutions. As the ions move they encounter frictional forces between themselves and the water molecules and other ions. The magnitude of these forces depends partly on the size of the ion so that, for example, ions of small size may move more quickly than larger ions. The ease with which an ion moves is defined in terms of the speed with which it moves in a unit electric field of 1 V/cm, which is referred to as the ionic mobility, u_+ for cations, u_- for anions. The way in which differences in ionic mobility may give rise to electric potentials may be seen by considering a system in which a salt is present at different concentrations

(C_1 and C_2) in two adjacent phases of a solution. As a result of the concentration difference between the phases, the ions of the solution will diffuse from the high-concentration to the low-concentration phase. However, if the cation, for example, moves more quickly than the anion, then an excess of positive charge will move into the dilute phase leaving an excess of negative charge in the concentrated phase. An electrical potential difference will therefore arise between the two phases. This potential is known as a *diffusion potential* and its equation was given by Planck in 1890:

$$E = \frac{u_+ - u_-}{u_+ + u_-} k \log_e (C_1/C_2), \tag{5}$$

where k is a constant defined in terms of the temperature, number of charges on the ions, and certain well-known physical constants.

The presence of a permeable membrane between the two solutions may have a very large effect on the diffusion potential if the membrane permeability is selective. Thus, if a membrane that is permeable to Na^+ but not to Cl^- is present at the interface between the two solutions, u_- becomes zero and the potential is then given by

$$E = k \log_e (C_1/C_2). \tag{6}$$

A fiftyfold ratio of concentration (which is quite commonly found in biological systems) would give a potential of about 0·1 V at normal temperatures. This equation (originally given by Nernst in 1889) describes a system in *equilibrium* in the thermodynamic sense: the free-energy levels are equal in the system and no energy is required for the maintenance of the potential difference. The idea that the cell-membrane potential may be an equilibrium potential of this kind became widely held in the first few decades of this century as a result of observations that the potentials in excitable cells are proportional to the logarithm of the ionic concentrations, as predicted by equ (6). The explanation that then seemed correct is that the cell normally contains a high concentration of K^+ ions and the external fluid contains a low concentration of K^+ ions, so that the potassium equilibrium potential given by equ (6) would be such that the cell should acquire a negative potential, as is in fact the case.

This suggestion requires that the cell should somehow acquire a high internal K concentration and it may seem difficult to reconcile this fact with the view that the system is an equilibrium state, since some energy must be used, or have been used, to establish the high internal K concentration. This problem might be solved in one of two ways. Either the energy required to establish the high concentration is provided by some 'initial' energy used to set the system up, or a continuous expenditure of energy must occur in order to 'maintain' a system that only approximates to a thermodynamic equilibrium. A state of the latter kind is distinguished by thermodynamicists as a *steady state*, and the controversy over this problem might be summarized in terms of the question whether the electrochemical state of a resting cell is an equilibrium state or a steady state. The importance of this question lies partly in the fact that the former state requires no further energy supply once the system has been set up, whereas the latter state requires a continuous supply of energy and this may account for an appreciable fraction of the total energy expenditure of the cell, particularly in small cells that have a relatively large surface membrane (compared to their volume) across which ionic exchange can take place.

For a long time (until about 1945) the view that the resting cell is an example of an electrochemical equilibrium state was widely held. It is interesting to consider why this was so. First, a good 'working model' existed, since a purely physico-chemical example of such an equilibrium had been described by Donnan in 1911 (although the theory had been previously developed by Gibbs in the 1870s). The *Gibbs–Donnan equilibrium*, as it is usually called, is established when a system of two aqueous phases is separated by a membrane that is completely impermeable to an ionic species present in unequal concentrations in the two phases. The inequality in the concentrations usually arises from the way in which the system is set up (the experimenter may literally put more of these ions in one solution than in the other). In living cells, it would be supposed, on this view, that the difference of concentration (usually thought to be formed by the high concentration of charged proteins and other large molecules that cannot pass through the cell membrane) is established during the processes of cell development. It is not possible here to give an account of the Gibbs–Donnan equilibrium but it will be sufficient for our purpose to note that, so far as permeant ions are concerned, the system eventually becomes an equilibrium system obeying equ (6).

The influence of this development in physical chemistry on the development of cell biophysics was very great indeed. It is an example of one of those ideas in science which have so much truth in them that, even when good evidence is found against the theory, the theory persists for a long time, particularly if the difference between the theory and its alternative is, on the one hand, important but not necessarily obvious (and this is frequently true of ideas that are distinguished largely in thermodynamic terms) and, on the other hand, numerically small and, perhaps, unimportant for many purposes. Thus, for many purposes, the potential predicted by assuming the cell to be in an equilibrium state with a membrane permeable to K^+ ions is so close to the observed potential that any differences may be either explained away or ignored.

The second reason why the 'equilibrium' view was popular is that for a long time the experimental evidence seemed to be strongly in favour of it. The potential observed in resting cells is proportional to the logarithm of the K concentration ratio. Moreover, it appeared to be completely independent of the concentrations of the other major extracellular ions, Na^+ and Cl^-. However, as has become very clear in more recent work, the lack of an obvious effect on the cell potential is not very good evidence for impermeability to a particular ion species.

The ready availability of radioactive isotopes since the Second World War has enabled the movements of ions across cell membranes to be studied more directly. In principle, although not necessarily in practice, the method involved is simple. Radioactive ions are placed on one side of the cell membrane and the speed with which they appear on the other side is measured. Using radioactive Na^+ ions it was possible to show that the cell membrane is not impermeable to sodium ions, as the equilibrium view required that it should be. However, it is also true that Na^+ ions are prevented from accumulating within the cell as K^+ ions do and this must be attributed to some process (usually called the sodium pump) that uses energy continuously. It is now clear in fact that both Na^+ and K^+ ions are actively pumped in opposite directions across the cell membrane, and the energetics and other properties of this pumping activity have been studied in great detail in recent years. The 'steady state' view of the resting cell is, therefore, well established. However, as indicated above, for some purposes it does not matter very much which of these two views is correct. Thus, the continued operation of the ion pumps is not *immediately* necessary for nerve and muscle cells to show electrical activity. It is not necessary, therefore, to discuss the developments in the study of ion pumps in order to describe the development of theories of excitation.

Early theories of excitation

A very important property of the excitation process that was known in 1900, although not established for all excitable cells (indeed, it is now known to be untrue for some cells), is that it is an all-or-nothing response to a graded stimulus. Thus, if a series of electric shocks of different intensities is applied to a nerve or muscle cell and the electrical or mechanical response is recorded, the cell fails to respond actively to stimulus strengths below a certain value. This value is called the *threshold value*. Stimuli above this value elicit an active response but this response is otherwise completely independent of the strength of the stimulus. This discovery has a number of important implications and, before discussing its significance in the development of theories of excitation, it will be of interest to briefly outline two of these implications in order to stress the importance of this property.

First, it is generally impossible for a single muscle fibre to give a graded response, since the muscle fibres respond to each nerve stimulus arriving at the nerve–muscle junction. Gradations in the strength of muscle contractions must therefore result from the *number* of muscle fibres that are fully active at a given time. This in turn means that gradations in muscle activity must result from variations in the numbers of nerve cells that are excited and the frequency with which the excitation occurs. This is one respect, though not the only one, in which the nervous system must operate in terms of quantal rather than continuous parameters.

Second, information cannot be transmitted in the nervous system in the way in which it is transmitted in many man-made transmission systems. In a telephone cable, for example, the signal current that flows along the transmission line is a continuously variable current that can be made to nearly reproduce, in electrical form, the variations in sound energy produced by the human voice. In the nervous system, however, all this information must be stored or carried by nerve fibres that can be in one of only two states: active or inactive. The continuous variations in the stimuli impinging on an organism must therefore be represented by a spatio-temporal pattern of unit excitations according to some code, much as messages can be transmitted in telephone cables by the dots and dashes of Morse code instead of by continuous variations in current strength. This fact has led to the frequent comparison of nervous mechanisms with those of digital computers whose components also operate in terms of two-state devices. However, this approach has not contributed very much to our understanding of the nervous system and has even been misleading. Thus, the nervous system uses a kind of frequency code in which the strength of a stimulus is represented in terms of the mean temporal frequency of occurrence of unit excitations, whereas digital computers operate in terms of strict binary representation of the variables being operated on. This is by no means a trivial difference: many important properties of the nervous system, particularly its ability to operate even when damaged, sharply distinguish it from a digital computer. This is not to say, of course, that we shall not eventually be able to make computers that accurately model the nervous system. But this approach has not contributed very much to the development of the subject up to the present time.

In addition to having important implications for neurophysiology, the existence of a discrete threshold for excitation was an important discovery in the development of nerve biophysics. Any adequate theory of excitation must account for the sudden explosive change in response that occurs when the stimulus exceeds threshold. It was natural, therefore, that early experimenters should concentrate first on the factors that determine the threshold for excitation. In the living body the stimuli that excite nerve cells are of various kinds. There are specialized nerve cells capable of responding to low levels of sound, light, heat, or mechanical energy. Until recently, however, it was difficult to control or measure these forms of energy accurately at the very low levels to which nerve cells will respond. Moreover, it is now known that the parts of excitable cells that are sensitive to these forms of energy are usually not those that are responsible for generating the all-or-nothing conducted response. In fact, it seems that all the diverse forms of energy are first transduced into electrical energy by the parts of the cells (usually the nerve-endings) that are especially sensitive to the stimulus concerned. We still know very little about this transduction process (we have no idea, for example, how a single light quantum is capable of initiating electrical activity in visual receptors). However, we do know that the electrical energy, if it is large enough, excites that part of the cell which generates the all-or-nothing response. It is this response (known as the *action potential*) that transmits the information to the rest of the nervous system. In some respects, therefore, it was almost a fortunate limitation that early experimental methods allowed accurate control of electrical stimuli but not of the more natural forms of stimuli. By using electrical stimuli, physiologists in the early part of this century could investigate the form of stimulus most closely concerned in the initiation of the action potential, although this could not have been much more than an intuition at the time and the justification for using electrical stimuli was usually based on the fact that only these stimuli could be easily and accurately controlled.

The simplest electrical stimulus is a constant-amplitude pulse characterized by its strength and duration, and early work concentrated largely on determining the threshold values of one of these parameters for different values of the other. The relation obtained between the threshold strength and duration turned out to be a relatively simple one and was consistent with the view, largely attributable to Nernst, that the excitable part of the cell is a polarizable membrane. If this membrane is represented by a capacitance and parallel resistance (as supposed by the transmission line model—see 'electric current flow in excitable cells', p. 209) then the observed strength–duration relation could easily be derived mathematically by assuming that the criterion for excitation is that a certain critical amount of electric charge should accumulate on the membrane. The strength–duration equation given by these assumptions was first derived by Lapicque in 1907 and Nernst in 1908. Although it is now known that several large additional factors have also to be taken into account, their effects fortuitously cancel in such a way that Lapicque's equation, and subsequent refinements of it by Hill and others in the 1930s, is closely obeyed. Thus, although there has been considerable argument concerning the precise theoretical significance of the criteria for excitation, the view that the membrane potential is the important factor soon became fairly well established. But what does this potential determine that could be responsible for the generation of the action potential? In order to see how this question was answered, we must first consider what was known about the action potential itself.

As described earlier (see 'electrochemical properties of excitable cells', p. 213) it was already known that a potential (the resting potential) exists between the inside and outside of the cell in the resting state. Also, it seemed clear that this potential is developed as a consequence of the large intracellular concentration of K^+ ions. It should be noted, however, that the experimental techniques available up to about 1940 did not allow direct measurements of this potential to be made. The measurements were usually made on the current that flows when the potential difference is short-circuited at one point of a cell by locally damaging the membrane. This current is called the *injury current* and its magnitude is proportional to (among other things) the membrane potential. This method of estimating the membrane potential introduces various kinds of error unless very careful precautions are taken. It is not therefore surprising in retrospect that the early attempts to measure the changes in membrane potential that occur during the action potential, ingenious though they were, failed to show one of its most important characteristics. The discovery of this characteristic (the 'overshoot'—see p. 221) was later to inaugurate the era of modern theories of excitation. However, its importance can best be appreciated in terms of its incompatibility with the theories of excitation based on indirect measurements, and I shall defer discussion of it until later. The indirect observations, however, did establish that after excitation by a suprathreshold stimulus a wave of potential change propagates along the cell. This wave was found to have the following characteristics.

(1) It consisted of a very short-lasting abolition of the resting potential.
(2) Its duration and magnitude are independent of the stimulus.
(3) Its magnitude does not change during propagation along the cell. As explained above (see 'electric current flow in excitable cells', p. 209) it would do so if a simple transmission-line mechanism operated.

Now the resting potential depends on the selectivity of the membrane permeability to K^+ ions. If the membrane were to become unselectively permeable to all ions, then the potential across it would be reduced to or near to zero, largely as a result of an inflow of Na^+ ions normally kept out of the cell by the relative impermeability of the membrane and the operation of the sodium pump. Overton, in 1904, showed in fact that excitable cells become inexcitable in the absence of sodium ions. It therefore seemed reasonable to suggest that the action potential is generated by a temporary loss of membrane selectivity, triggered in some unknown way by the deflection of the membrane potential beyond the threshold value. This theory was proposed by Bernstein in 1913 and it easily accounted for the facts as they were then known. Moreover, by involving transmission-line theory it could also account for propagation. The disappearance of the cell potential at one point would cause a current to flow between this point and those parts of the cell that were still polarized, and this current flow would deflect the potential of the polarized areas in the direction of the threshold. Provided that this local circuit current flow were strong enough, the polarized areas would also become active. The mechanism of the initial flow of current would be similar to that described by linear transmission theory: only when the threshold voltage was reached would an additional mechanism come into operation.

It is of great importance in the development of any field of science to evolve plausible theories. But it is quite as important to develop experimental methods for testing them. The fact that the electrical propagation theory could be stated in a plausible way was insufficient by itself to settle the issue between electrical and chemical theories. The crucial quantitative test of the electrical theory came in 1937 when Hodgkin made measurements of the electric current flow generated by the action potential. Hodgkin's technique was to make a length of nerve cold enough to fail to conduct an impulse. He then showed that if the length of cold nerve was great enough the action potential initiated on one side of the cold block region failed to generate enough current to excite the regions of the cell beyond the cold block. However, it was still possible to record the current beyond the blocked region and to show that it could summate with applied currents to initiate an action potential. Calculations of the magnitude of current that would have flowed in the absence of the block showed that the current is considerably greater than the minimum required to ensure propagation. It was therefore no longer possible to hold that the electrical theory of transmission is inadequate. However, there still remained the problem of the coupling between membrane potential changes produced by the local circuit current flow and the permeability changes responsible for the explosive depolarization.

The modern theory of excitation

This problem might still remain unsolved but for an extremely fortunate natural occurrence discovered by Young in 1937. The speed of propagation of the nerve impulse is a function of the size of the nerve, and increases as the size increases. In order to achieve fast conduction for the purpose of activating some muscles very quickly, some invertebrates have developed exceedingly large nerve fibres. The squid, for example, has fibres up to 1 mm in diameter. Young's discovery did not go unnoticed (as had the same discovery long before) and by 1939 physiologists in England (Hodgkin and Huxley) and the U.S.A. (Cole and Curtis) had achieved the technical feat of inserting recording electrodes inside these giant nerve fibres. This enabled the true membrane potential to be recorded for the first time. The results differed from the earlier indirect results in one very important respect: during activity the membrane potential does not simply become zero. It swings over to a positive value (see Fig. 10.2). This phenomenon, known as the *overshoot*, could not be accounted for by Bernstein's theory. Interestingly enough, the possible existence of an overshoot had been noticed by some workers using indirect methods in the nineteenth century, long before Bernstein's theory was formulated. The significance of the observations was not appreciated (probably because there existed no theoretical framework within which to appreciate their significance) and the discovery was completely forgotten until recently. Moreover, the possible errors in the indirect methods and the appeal of Bernstein's theory were probably great enough to deter anyone else who may have obtained indirect evidence for an overshoot from publishing it. The internal methods were so direct that the possibility of an error large enough to account for the overshoot could be virtually eliminated.

There was time only for preliminary reports of these results to be published before the onset of the Second World War. It was not, therefore, until 1949 that a new theory to replace Bernstein's was properly formulated, and only in 1952 was a complete mathematical formulation of the modern theory of nerve excitation and conduction given.

Hodgkin and Katz's work in 1949 was based on the observation that nerves become inexcitable in solutions that contain no Na^+ ions. By measuring intracellular action potentials in various concentrations of extracellular sodium ions they were able to show that the overshoot potential nearly obeys the relation

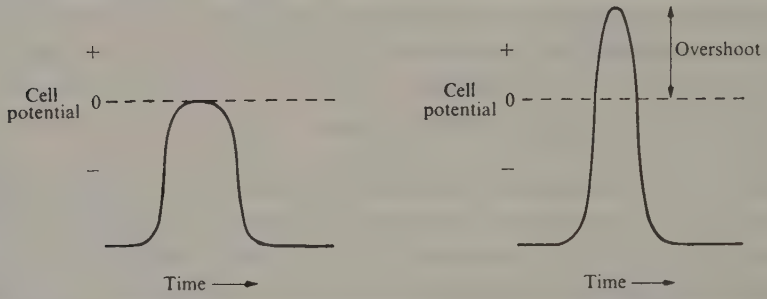


FIG. 10.2. *Left:* potential changes during activity predicted by Bernstein's theory. The membrane simply loses its potential for a period of time as a result of a loss of selective permeability. *Right:* potential changes actually recorded by direct methods using electrodes placed inside nerve cells. Note the presence of a marked overshoot that cannot be accounted for by Bernstein's theory. In the modern theory of excitation this effect is attributed to a *change* rather than to a loss of membrane selectivity.

$$E_{\text{overshoot}} = k \ln \frac{[\text{Na}]_o}{[\text{Na}]_i}, \qquad (7)$$

which gives the potential that would occur if the cell membrane were highly selectively permeable to Na^+ ions: cf. eqn (6) above. Hodgkin and Katz therefore concluded that instead of becoming unselective in its permeability, as in Bernstein's model, the membrane changes its selectivity from K^+ to Na^+ as a result of a large increase in permeability to Na^+ ions.

This conclusion was important but it still did not answer the question of how changes in membrane potential can produce such large changes in membrane selectivity. At the molecular level this question is still largely unanswered, but considerable progress in the direction of quantifying the factors involved was made possible by the invention by Cole and Marmont in 1949 of an ingenious method for controlling the cell-membrane potential. This method, which is known as the 'voltage clamp' technique, uses electronic feedback to feed current through the cell membrane in such a manner as to hold the membrane potential at whatever value is chosen by the experimenter. By measuring the currents required to hold the membrane potential at a whole range of different potentials, and by identifying the ions carring these currents across the membrane it became possible to measure the influence of the membrane potential on the ionic current flow, and hence on the membrane permeability. The results of an extensive analysis of this kind by Hodgkin and Huxley in 1952 led to their formulation of the modern theory of excitation.

It is not possible here to give even a simplified account of the Hodgkin–Huxley theory in mathematical terms. However, by giving a fairly concrete physical interpretation to the theory (which, it should be emphasized, forms only one of many possible physical interpretations) a useful non-mathematical account may be given.

The existence of an electric potential difference across the membrane means that the membrane itself is subjected to an electric field and, since the membrane is known from electron microscope and other studies to be extremely thin (of the order of 100 Å), the field is likely to be very intense (of the order of 100 000 V/cm). Some of the molecules that form the membrane structure are charged, and the position of these molecules may therefore be determined by the strength and direction of the field. Suppose that some of these charged groups are placed in the membrane channels that are thought to conduct the ionic currents and that, when the field strength is reduced or reversed (as will happen when the membrane is depolarized by current flow), the groups move into a new position. One of these positions may impede the movement of ions through the channels more than the other. The charged groups will then act as field-sensitive gates controlling the membrane permeability to ions. The general behaviour of such a system can be described mathematically, although at present not all of the constants can be derived from purely theoretical considerations. Hodgkin and Huxley found in 1952 that the ionic current and its dependence on the membrane potential could be adequately described by postulating the existence of three kinds of gating mechanism of this general type. One of these mechanisms, for example, controls the flow of Na^+ ions. These 'gates' open when the membrane is depolarized to a certain extent, so allowing the Na^+ permeability of the membrane to increase. Na^+ ions then flow into the cell, thus further reducing the internal negative potential. As a consequence even more of the sodium gates open and the current flow increases. In this way a regenerative mechanism is produced. Once this mechanism is activated by depolarization it will continue to depolarize the membrane of its own accord. In this way we can easily account for the presence of the threshold and for the all-or-nothing behaviour of the action potential referred to above (see 'early theories of excitation', p. 216). The equations describing the three gating mechanisms were sufficient to describe accurately the ionic current as a function of membrane potential and time, and we may represent them by a general function:

$$\mathrm{I_i} = \mathrm{F}(V, t). \qquad (8)$$

If this function is used in place of GV in the transmission-line equation described earlier (equ (4)) it becomes possible at last to amalgamate the electrochemical and transmission-line properties of the nerve cell into one mathematical theory. The resulting equations cannot be solved by the methods that are available for the linear equations. The calculations are in fact very tedious if done by hand and the development of fast digital computers in the past two decades has greatly helped in developing the modern theory. Nevertheless, many of the major results were obtained by hand computation by Hodgkin and Huxley in their 1952 papers.

Fig. 10.3 shows a schematized version of the theory and shows the variation in membrane potential, permeability (expressed in terms of the sodium and potassium conductances), and ionic currents with distance along a nerve fibre during propagation of an action potential. The sequence of changes may be roughly divided as follows.

(1) An initial roughly exponential phase of depolarization, which is produced by spread of current from the advancing wave. This corresponds to the time during which the ionic current is negligible, so that the mechanisms described above have not yet been activated. This phase also corresponds to a time in which the system is behaving very much like a linear transmission system.

(2) The opening-up of the sodium gates consequent upon the depolarization becoming large enough to initiate the regenerative flow of sodium ions into the cell.

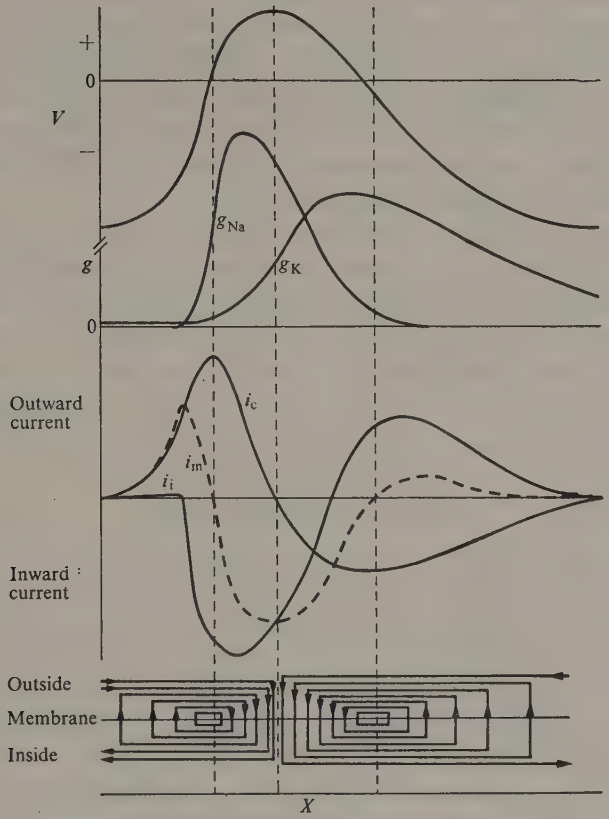


FIG. 10.3. Modern theory of nerve excitation. Curves show changes of conductance and ionic currents flowing across membrane during propagated action potential. Interrupted lines indicate corresponding points on curves. Lowest diagram shows local circuit currents flowing between different regions of membrane. First stage of depolarization, 'foot' of action potential, occurs before any appreciable rise in g_{Na} or i_i occurs. During this time the membrane current is generated by other areas of membrane where large conductance changes have already occurred. This local circuit depolarization then triggers changes in conductance that generate the rest of the action potential. From Noble (*Physiol. Rev.* **46**, 1–50 (1966)).

(3) The recovery phase, during which the sodium channels close down (this is attributable to a second gating mechanism that operates in the opposite direction to the first) and the potassium channels open up (this is attributable to the third gating mechanism that controls channels specific to K+ ions).

Thus, the membrane permeability automatically swings from being K^+-selective (resting) to Na^+-selective (active) and back again to K^+-selective (recovery). The system is automatic in the sense that it proceeds without further help once the threshold for opening the sodium gates has been exceeded and also in the sense that the energy for driving the gating mechanisms comes from the electric field itself. As described above (see 'electrochemical properties of excitable cells', p. 213) the energy for the generation of the electric field comes from the differences in ion concentrations across the membrane, and these ultimately depend on the activity of the ionic pumps in the membrane.

This, then, is the picture that emerges after half a century of work in this field. The problems now concern the molecular mechanisms in cell membranes, about which we as yet know very little. However, there is a sense in which cell biophysics has almost completed its task in the problem of nerve excitation. The next phase should be molecular-biological.

Selected References

BERNSTEIN, J. (1902) *Pfluger's Arch. ges. Physiol.* **92**, 521–62.

HERMANN, L. (1905) *Pfluger's Arch. ges. Physiol.* **109**, 95.

HILL, A. V. (1932) *Chemical wave transmissions in nerves.* Cambridge University Press.

HODGKIN, A. L. (1937) *J. Physiol., Lond.* **90**, 183–232.

HODGKIN, A. L. and HUXLEY, A. J. (1952) *J. Physiol., Lond.* **117**, 500–14.

HODGKIN, A. L. and KATZ, B. (1949) *J. Physiol., Lond.* **108**, 37–77.

LAPICQUE, L. (1926) *L'excitabilité en fonction du temps.* Paris.

OSTERHOUT, W. J. V. and HILL, S. E. (1930) *J. gen. Physiol.* **13**, 547.

OVERTON, E. (1902) *Pfluger's Arch. ges. Physiol.* **92**, 346–86.

YOUNG, J. Z. (1936) *Q. Jl microsc. Sci.* **78**, 367.

11 *The Viruses*

It is tempting, by hindsight, to see in ancient writings a knowledge of infective processes. The classically minded can cite Varro, the literary, Bocaccio, or Shakespeare in such passages as

> Take thou some new infection to thy eye,
> And the ranke poyson of the old will die.

Scientists cite Fracastoro, Kircher, Bradley, and many others who were less influential. Up to the time of Pasteur, only a minority held beliefs comparable to those generally accepted today: Linnaeus, for example, thought that infectious diseases were caused by acarids and worms. The absence of a rational theory of infection did not interfere with the prosecution of vigorous and effective public health measures by people like Ozanam and Nightingale, nor with the detailed documentation of epidemics by Hecker. After Pasteur opinion swung steadily in favour of the idea that infections had a microbial origin in plants, animals, and people. The recognition of infective states in bacteria had to wait till 1915.

The distinction between viruses and other infective agents depended on progress in microscope design towards the end of the nineteenth century and on the manufacture of filters from kieselguhr taken from the mine belonging to Herr Berkefeld. The word *virus* is ancient but was used imprecisely although Celsus, probably inadvertently, used the world *virus* when speaking of rabies but *venenum* when speaking of snake bite. Pasteur, who postulated infective agents too small to be seen with a microscope, did not distinguish between viruses and microbes. Thus, discussing the invisible infective agent of rabies, he said '*tout virus est un microbe*'. Negative evidence, that is to say the microscopic clarity of infective fluids, was complemented at the end of the century when Loeffler and Frosch found that the symptoms of foot-and-mouth disease were transmitted by fluid that had passed through a bacterium-tight filter and argued convincingly from subsequent passage experiments that it multiplied in the host. Beijerinck found this with tobacco mosaic virus at about the same time. Other similar agents were soon found and they were grouped together as filter-passing, or filterable, viruses. With usage, the simple term virus became restricted to them.

Viruses were therefore characterized by the use of two criteria that depended on their small size. This led to much speculation about their fluid or non-particulate nature. By hindsight this is totally confused and it is hard to believe that it did not seem very confused, or even nonsensical, to contemporaries with any chemical or biochemical training. By the beginning of this century, when the virus category was recognized, there was almost universal agreement that Democritus and Dalton were right: matter was made from atoms and so everything was particulate. The only uncertainty concerned the sizes of the particles.

The use of particle size as the criterion for putting an infective agent into the virus category led to the inclusion in that category of several agents that were later excluded. These are now grouped uneasily together as the Mycoplasmataceae or Pleuropneumonia-like organisms, PPLO for short. They cause 'virus' pneumonia, agalactia, and some of the ornithoses. Non-pathogenic members of the group were found by Laidlaw in sewage, and very interesting early biochemical work was done on 'viruses' when their enzyme activities were studied. They are now excluded because they can multiply in culture media and do not need an infectible host cell. The present-day criteria for a virus are that it should not only be small and cause recognizable symptoms in some host, but that it should fail to multiply in any inert medium. These criteria are applied somewhat loosely so that a more practical definition of a virus would be that it is 'any infective agent that is not more conveniently classified as something else'. Furthermore, the requirement that it should fail to multiply in an inert medium has all the well-known defects of a negative criterion and it is rather too inclusive; several infective agents that no one wishes to classify as viruses have so far not been cultivated without a host.

Partly because of this early grouping together of agents that we now assign to separate categories, and partly because of the metaphysical superstructure that tended to accrete round the concepts 'chemical substance' and 'organism', an experimentally useful understanding of the significance of the need for a host in virus reproduction was reached very slowly. As early as 1914 Sanfelice suggested what is essentially the present-day interpretation of the nature of virus infection, and Bordet, in conflict with Twort and d'Herelle, held a similar view of the nature of bacteriophage infection. Mulvania suggested that the various attributes of life could exist independently and that the viruses possessed only some of them and so, although having life-like properties, could not be regarded as being fully alive. But most of those who studied viruses before the mid 1930s, were convinced that they differed from other microorganisms in no significant respect other than their small size. I remember many heated discussions with Gye, whose early and vehement advocacy of the role of viruses in causing cancer is now insufficiently recognized. He had studied chemistry to an extent that was most unusual with virus workers at that time; this may have given him a rather rigid approach for he was unable to visualize an agent of determinable structure and composition producing the effects with which he was familiar in infection. It is however interesting that he had a 'hunch' that viruses worked through the nucleoproteins of the cell although at that date (1932) nucleoproteins were a plausible biochemical myth; they had not been studied in solution but were simply postulated.

A biochemical outlook was however spreading into microbiology. Stephenson started a vigorous group in Cambridge investigating the metabolism of bacteria. Stimulated to some extent by this, work started on the metabolism of some members of the PPLO group, on the biochemical properties of some viruses, and on the processes by which bacteriophages become attached to the host. Because of lack of interest on the part of grant-giving bodies, this work had to stop; it restarted elsewhere a decade later.

It is hard for those who did not experience the conditions of work 30 to 40 years ago to understand its pleasures and difficulties. There was as much literature to read as there is now because we tried to know something about the whole of the subject and not just a facet of it; but it was pleasant to be able to be sure that one knew everything that had been published on one's own particular facet. And most subjects, biochemistry among them, were so fluid that novel observations could lead logically to an extensive range of possible interpretations. On the other hand, even those with an academic record unsullied by anything as low as a good second class degree, now a certain qualification for a research job, often had difficulty finding support for projects without direct agricultural or medical application.

The change in equipment has more specific relevance to virus research. Filtration methods have improved, but not out of all recognition. Centrifuges, on the other hand, have altered radically. Analytical centrifuges able to sediment the larger proteins were in use in the late twenties but they were extremely expensive and not suitable for handling the volumes of fluid needed for the preliminary separation of viruses for serological or infectivity measurements. During the thirties there was rapid technological improvement. The air-supported and driven spinning-tops that Henriot and Huguenard had introduced into physics laboratories had a brief vogue but they had small capacity. They were the forerunners of the air-driven centrifuges, made by Beams and others, which dominated the field during the forties. Bechold, who had already contributed notably to virus research by introducing nitrocellulose membranes of known porosity, studied the possibility of increasing the speed of electrically driven centrifuges of the conventional bucket type. By 1935 it was possible to centrifuge 50 ml at 16 000 r.p.m. but people travelled half across Europe to get access to such machines. They soon became more widespread. I tried to persuade several British manufacturers to make centrifuges of this type, but was unable to convince them that a world-wide demand was imminent however small the market appeared at the moment. Spinco owes much to my inadequacy as an advocate.

High-speed centrifugation was, and still is, the pre-eminent method for purifying viruses. Schlesinger, who had worked with Bechold and was welcomed by Gye when the Nazis forced him to leave Germany, used an electrically-driven bucket centrifuge in the purification of a bacteriophage and separated it in '*mit freiem Auge sichtbaren Mengen*'. The preparation was not only visible but there was enough of it for analysis; the presence of 3·7 per cent of phosphorus, and the Feulgen reaction given by the preparation, suggest that it contained deoxyribonucleic acid and was about as highly purified as most preparations made many years later. Bechold and Schlesinger were also able to sediment tobacco mosaic virus centrifugally, but the accelerations they could attain were too small for this to be a satisfactory preparatory method with such a small virus.

There is little scope for variations on the basic theme of centrifugation; machines have become easier to use and freer from vibration. Improvements have been made in the optical arrangements for locating a boundary when measuring sedimentation rate. Temperature control and the diminution of frictional heating of the rotor have made convection less disturbing. It was made still less disturbing twenty-five years ago when Pickles introduced the principle of centrifuging in a density gradient and this has also made it easier to separate mixtures of particles of differing density. Other changes have concerned matters of detail only.

The impact of high-speed centrifugation on virus research was immediate and complete. Nearly every institute in which viruses were studied installed a machine though not all of them were regularly used. The electron microscope, on the other hand, was adopted more slowly. There are several reasons for this. The early instruments were more expensive and temperamental than centrifuges and, even with skilled staff, suffered frequent breakdowns. Furthermore, it was seldom possible at first to be able to resolve virus particles in tissues or crude extracts so that a microscope did not become useful until equipment for virus purification had already been installed. The possibility of electron-microscopy was demonstrated in 1931 in several different institutes, but the magnification was only ×400. Two years later Ruska achieved ×12 000 and so surpassed the best quartz-lensed ultra-violet microscopes. These early electron microscopes were used to study electron-emitting surfaces and the edges of metal and mineral particles; they were not used to study viruses until 1939. The reasons for the delay are not obvious because the work that was then done with viruses did not necessitate cutting thin sections; it was the slow development of that technique, and the slow development of microscopes with sufficiently powerful electron beams to penetrate sections, that delayed electronmicroscopic study of tissues. In 1940 there was a burst of activity on both sides of the Atlantic. The conclusions about the size and shape of tobacco mosaic virus and tomato bushy stunt virus, that had been based on X-ray crystallographic and physico-chemical considerations, were confirmed and the tadpole shape of a bacterial virus was discovered.

Steady improvement in the quality of the electron-focussing devices made great increases in the effective magnification possible; even a magnification of ×400 000 does not approach the limit imposed by wave theory though it may be approaching limits set by some other factors. That magnification is sufficient however to reveal details of surface and internal structure. During the forties and fifties many contradictory claims were made, but electronmicrograms have only recently been able to supplement the information gained by crystallographic methods. The difficulty arises because of the small electron-stopping power of organic matter: it is the precise inverse of the difficulty encountered in studying the surfaces of impenetrable pieces of metal or mineral. The latter difficulty was partly circumvented in 1942 when Müller evaporated a film of metal from a point source to one side of the specimen so that it cast ‘shadows’ across the surface and brought out detail. This method was soon adapted for use with viruses and gave such enhanced contrast that clearer photographs could be made at greater magnification. Contrast was also increased by various staining reactions. Viruses had at one time been so heavily loaded with stains as to be visible by light microscopy, and some aspects of the structure of bacteria had been studied by ‘negative staining’ with such dyes as nigrosine. Both these processes inspired the use of metals deposited from solution to enhance the contrast in virus particles and an enormous amount of information about internal and external structure has been gained. Not unexpectedly, there are complications. It is not always clear which parts of the final pattern are derived from the back, the front, or the middle of the particle; nor is it certain that regions of dense metal deposition are holes, as the nigrosine analogy would suggest, rather than regions of special chemical reactivity. But improvements in method and interpretation are now rapid and electronmicroscopy has become the dominant technique for investigating virus structure.

When bacteriologists speak of a culture being pure they mean that it is free from other infective agents recognizable in the hosts or on the media that they are using: they are untroubled by the presence of residual culture medium, agar, or cotton wool, and they often do not record the chemical composition of the bacterial mass. This was at first the attitude of mind of virus workers also, but the scale of virus research increased at a time when biochemical ways of thinking were spreading, and Beijerink’s concept of the ‘infective fluid’, irrational as it may have been in principle, tended to encourage the idea that viruses were entities that might usefully be purified and analysed chemically. In theory it could be argued that this was an unlikely project. Centrifugal and filtration evidence about the sizes of viruses, taken in conjunction with their dilution end points for infectivity, suggested that even as infective a starting material as the sap from leaves infected with tobacco mosaic virus would, if one particle could cause an infection, contain only a few milligrams of virus per ton.

As already mentioned, a bacterial virus was purified by sedimenting it in a high-speed centrifuge. The purification of the first plant viruses depended simply on the application of standard methods that had been used in protein chemistry for fifty years. During the twenties and early thirties it was commonly assumed that viruses were, or contained, proteins; it is therefore surprising that successful purification was so long delayed. Several reasons can be suggested: plant pathologists were at first not familiar with protein chemistry; they sometimes tried to purify viruses that even now present serious difficulties; they often tried to use methods that were recent innovations but that have not remained popular; and they could not believe that the product they sought could be a prominent component of sap. The last reason is probably the most significant, especially for those who worked on tobacco mosaic virus; on many occasions material must have been discarded as 'impurity' when in fact it contained a much greater proportion of virus than the supposedly purified preparation.

The agents causing tobacco mosaic and foot and mouth disease were the first whose status as viruses in the modern sense was established. Tobacco mosaic virus was also the first to be associated with a characteristic physical property for, in 1933, Takahashi and Rawlins observed flow birefringence in sap from infected, but not healthy, plants and suggested that this was caused by the presence of rod-shaped virus particles. Two years later the crystallization of the virus was claimed by Stanley. The claim was at first treated with reserve by most virus workers —a reasonable attitude for nearly all the specific statements he then made about the nature of the virus have not been substantiated. Stanley's paper is widely quoted by people who do not seem to have read it. Preparations with properties more closely resembling those now ascribed to the virus were made by Bawden and Pirie in 1936. These showed flow birefringence dramatically and, at a critical concentration, became liquid crystalline. In the next two years we made liquid crystalline preparations of five other plant viruses, and fully crystalline preparations of tomato bushy stunt virus. One of the group of tobacco necrosis viruses was also crystallized.

Thoughtful scientists realized immediately that crystallization, or the separation of virus preparations in forms with characteristic physical properties, adds nothing to our basic understanding of the nature of the infective process or of the processes by which viruses multiply in a host. It does not even supply significantly better evidence for homogeneity than was being supplied by such processes as ultracentrifugation and electrophoresis. Nevertheless, crystallinity caught the imagination of the less thoughtful and evoked a flood of comment on such themes as 'crystalline life' and 'living molecules' that was reminiscent of the older illusions about non-particulate viruses. The perpetrators of all this had clearly never paused to consider what exactly they understood by concepts such as life, molecule, particle, and crystallinity, which they were bandying about. The preparation of many viruses in crystalline or orientated forms did however make them amenable to study by X-ray crystallography and this has been, and is being, of immense value in the determination of virus structure.

All these plant viruses were ribonucleoproteins; many years elapsed before the presence of lipids, polyamines, and non-ionically bound metal, was recognized in some of them. So far no plant virus has been found that contains deoxyribonucleic acid whereas this nucleic acid was present in all the preparations of animal and bacterial viruses that were at first made. Later work showed that in some of them the deoxyribonucleic acid was a contaminant and that these viruses may contain either type of nucleic acid, but no virus is known that contains both together. The composition of many animal and bacterial viruses is much more complex than that of most plant viruses: lipids are often present and, after fission by methods that would not be expected to cause protein dissociation, several distinguishable proteins are found in some of them. These apparent differences between plant viruses and the others are very interesting and may be significant, but it would be premature to put much weight on them. It is now obvious that no rigid distinction can be drawn between plant and animal viruses because some viruses are able to multiply and to cause disease symptoms in both types of host. The larger viruses tend to be the ones that are chemically the more complex and the plant viruses so far isolated, perhaps because extraction is not so easy from plant as from animal or bacterial cells, tend to be small. Furthermore, although nucleic acids are probably an integral and essential part of the structure, the status of some of the other virus components is less clear. Thus tobacco mosaic virus, as normally prepared, contains host protein, ribonuclease and some other enzymes, and traces of host nucleic acids of both types. The virus is so robust that, by various treatments, these components can be removed, though not without changing the physical properties of the virus significantly. With less stable viruses this is not possible and many, for example herpes and influenza, carry with them material that varies according to the host tissue used in the cultivation of the virus.

Now, after thirty years work on purification, no virus is known in which nucleoprotein is not the major component, nor is one known that has an array of enzymes similar to those that carry on the energy exchanges in conventional cells. Some viruses do however have enzymes that facilitate their penetration of the host cell wall. Some, for example scrapie virus, seem to have properties far removed from those of the familiar nucleoproteins, but no such aberrant virus has as yet been purified to such an extent that chemically significant statements can be made about it.

At the beginning of the century the usual assumption was that biochemical specificity depended on proteins. Then, in the middle of the period under review, the specific potentialities of polysaccharides were recognized in bacteriology and vigorous efforts were made to involve them elsewhere. The significance of the nucleic acid component of viruses was, at first, disputed, especially in the U.S.A. where purified preparations were often referred to as 'virus protein', a usage I condemned at the time as either redundant or over-precise. The text books said that nucleic acids were tetranucleotides. People who had not read the relevant papers carefully enough to see how flimsy the evidence was on which this hypothesis rested, accepted the hypothesis and with it the corrollary that the number of possible permutations in nucleic acids was too small for them to be vehicles of specificity. Those of us, on the other hand, who handled nucleic acids made by gentle methods, were familiar with their indiffusibility, and with the viscosity and small osmotic pressure of their solutions. We therefore never took the hypothesis seriously and were as ready to ascribe specificity to the nucleic acid as to the protein. Pfankuch was perhaps the first to state this categorically, but the point was not accepted until Avery and his colleagues showed in 1944 that nucleic acid was the agent that transformed 'rough' strains of pneumococcus into 'smooth' strains. The pneumococcus transformed our opinions about nucleic acids as radically as nucleic acids transformed the pneumococcus.

Interest in the nucleic acids increased still more when Hershey found that most of the protein in a bacteriophage did not enter the infected bacterial cell. The bacteriophage used was however so large that the amount of protein that does enter the cell is of the same order as the amount of protein in the small plant viruses; the conclusion to be drawn was therefore uncertain. In 1956 infective preparations of nucleic acid that were essentially free from protein were made from tobacco mosaic virus. This completed the establishment of the period of nucleic acid obsession. From having been unreasonably neglected in the thirties, nucleic acids are now in danger of being unreasonably invoked in almost every biological activity involving specificity. Experience in the heyday of proteins and polysaccharides suggests that attempts at what might be called 'biochemical monotheism' are not likely to be successful. If it should turn out that the scrapie virus is not a nucleoprotein, those who have passed through all these phases will be interested but not astounded.

The smaller viruses are now commonly thought of as nucleic acid knit together by protein and some other substances into a structure that is sufficiently robust to resist digestion and other forms of inactivation within the cell. By a mechanism about which there is little information, the structure is dismantled in the host and the liberated nucleic acid stimulates the synthetic mechanisms of the host to make virus-type nucleic acid and protein instead of the types they make normally.† In essence this picture fits the facts satisfactorily. It makes sense of the geometrical simplicity and compactness of many viruses, for these forms are likely to give the necessary stability. Some accounts of virus architecture imply that it is *because* they are viruses that the regular form is adopted: I suggest rather that *unless* they have one of these forms they are unlikely to be able to be viruses. Specificity in the dismantling process would do much to explain the peculiarities of host range; thus some viruses will infect only some varieties within a species although they are able to infect species in several different families, and others, for example poliomyelitis virus, have had their host range extended by techniques, such as growth in tissue culture, that are thought to change the protein component. Furthermore, with plant viruses at any rate, the physiological state of the host greatly affects its susceptibility to infection. One factor responsible for this may be the inactivation of nucleic acid, in the liberated and vulnerable state, by metabolites and other variable components of the host. The processes of dismantling and nucleic acid inactivation *in vivo* deserve very much more attention than they are getting at present.

This picture of virus reproduction or replication goes far towards resolving the old conflict between the endogenous and exogenous theories of virus, and especially bacteriophage, origin that used to be disputed among advocates of the Bordet and d'Herelle points of view. As recently as 1953 these were still having repercussions. The exogenous school thought of viruses as organisms that found in a host a culture medium on which they could grow; the endogenous school regarded the host as capable of engendering a new transmissible entity regularly as the result of some form of metabolic or environmental insult. Both outlooks are largely reconciled if we think of a virus as something that has to use the host synthetic machinery in an active manner, rather than as something that can use the host as a supplier merely of raw material. This means that only those viruses that find in a cell a mechanism adapted to the

† For an account of the mechanism of nucleic acid and protein synthesis, see Chapter 7.

reproduction of virus structure, can use that cell as a host. As a corollary of this, the host may control not only the extent but also the type of synthesis that takes place as a result of virus intrusion. Consequently we would expect the 'spontaneous' appearance of a virus in an uninfected host to be a possible but rare event, thus satisfying Bordet's claims, while at the same time expecting the host to exert much more influence over the progress of virus multiplication than would be reasonable on the d'Herelle picture.

Complex as this outline still seems to many virus workers, the actual position is still more complex. It has been known for many years, but the fact has been both disputed and neglected, that virus infection can lead to the synthesis of a range of products that resemble the virus in many ways but that do not seem to be infective. At the very end of the period under review, the study of what is now called 'satellitism' has partly clarified this phenomenon. In plants, animals, and possibly bacteria, the multiplication of one agent may depend on the presence of another. This phenomenon is not comparable to transmission by a vector such as an insect, mite, nematode, or fungus, it is equally well manifested when the agents are transmitted mechanically. Often, one component of the mixture that occurs in an infected host is an infective agent in its own right, whereas another cannot multiply unless the first is present. Sometimes, however, there is a state of affairs comparable to symbiosis in which each component helps the other to multiply. The form that this 'help' takes is still uncertain but the recognition of the phenomenon will do much to dispel any complacent tendency to think that, even if the details of virus infection remain to be clarified, the fundamental mechanism is now understood.

Selected References

Bawden, F. C. (1939, 1942, 1950). *Plant viruses and virus diseases.* Chronica Botanica Company

Fildes, P. and van Heyningen, W. E. (1953) *The nature of virus multiplication.* Cambridge University Press.

Kassanis, B. (1968) Satellitism and related phenomena in plant and animal viruses. *Adv. Virus Res.* **13**, 147.

Lwoff, A. (1966) Interaction among virus, cell, and organism. *Science* **152**, 1216.

Medical Research Council (1930) *A system of bacteriology*, Vol. 7, *Viruses.* H.M.S.O.

Pirie, N. W. (1962). The principles of classification illustrated by the problem of virus classification. *Perspect. Biol. Med.* **5**, 446

—— (1962) Patterns of assumption about large molecules. *Archs. Biochem. Biophys.* Suppl. **1**, 21.

Wolstenholme, G. E. W. and Millar, E. C. P. (editors) (1957) *The nature of viruses.* J. & A. Churchill.

12 *Ethology*

I

In the last half century we have witnessed the development, or rather the revival, of a science of animal behaviour that is now widely known as ethology. Like so many other scientists, ethologists find it difficult to characterize their science. They agree that (as expressed in the word ethology) the starting point of their studies is an observable—'habits', in the sense of movements—and it is relevant to an understanding of their approach that they are biologists. I myself like to define ethology as the biological study of behaviour, a formulation that mentions both the observable phenomenon and the method of study.

One of the forerunners, though because of lack of immediate followers not a founder of ethology, was Darwin (1859, 1872). With remarkable foresight he realized that if his theory were to explain evolution of animal species by means of natural selection he had to apply it to all properties of animals, whether 'structural' or 'functional', and therefore could not ignore behaviour. His work contains much concrete material that we would now call ethological. While it was at that time impossible to know much about the 'machinery' of behaviour, Darwin's procedure could be characterized by saying that he treated behaviour patterns as *organs*—as components of an animal's equipment for survival.

The ethological aspects of Darwin's work were not followed up at once. Biologists after Darwin concentrated until well into the twentieth century on the elaboration and consolidation of a picture of phylogeny, i.e. of the course evolution must have taken, and for this they selected the most obvious observables: structural characteristics. Thus for a long time zoology at least concentrated largely on comparative anatomy.

Now and then isolated starts towards a biological science of behaviour were made by individuals, without, however, setting off an immediate, continuous development. Neither Morgan's work on instinct and habit (1896), nor Whitman's sophisticated comparative studies of simple behaviour patterns (1919), nor Jenning's studies of the behaviour of unicellular organisms (1923), to mention a few outstanding examples, really caught on. The explanation of this failure is undoubtedly complex: the existing branches of biology, to which genetics had in the meantime been added, attracted the available talent; there were relatively few biologists, and their science did not have much of a status even in scientific circles; behavioural phenomena were too complex to yield easily to exact description, let alone to experimental analysis; above all, perhaps, religious attitudes hampered scientific analysis of animal behaviour, so uncomfortably reminiscent of our own.

In the first quarter of the century, however, two biologists began to produce work that had a more immediate impact. In Britain Huxley (1914, 1923) began to redevelop the study of bird behaviour, particularly mating behaviour, with a view to understanding its evolution by natural selection. He argued that success in reproduction depends on epigamic displays and structures combined, and that both were parts of signalling systems in which each sex partner in a pair bond exerted the selection pressures that favour increased effectiveness as signals. In Germany Heinroth (e.g. 1911, 1928), in a series of deceptively simple but in fact highly sophisticated publications, applied the methods of comparative anatomy to behaviour patterns, mainly of birds, stressing the fact that many behaviour patterns are typical of species, genera, and even families and larger taxonomic groups. He also demonstrated the fact that many of these behaviour patterns were largely 'innate', in the sense of being not or hardly modifiable by variations in the environment in which individuals grow up.

It was perhaps no accident that both men worked mainly on the behaviour of birds. Birds are on the one hand sufficiently highly developed to make their behaviour seem a special, challenging phenomenon (unlike the seemingly more machine-like movements of, for example, a starfish) but on the other hand are less likely than, for instance, monkeys to be identified with ourselves.

The impact of their work was at first confined to the more scientifically-inclined bird watchers. A steadily swelling stream of papers on bird behaviour began to appear, of which some were of a very high standard (e.g. Verwey 1930). But the real start of a more comprehensive attack was made by one individual, Lorenz of Vienna, whom Huxley was later (1963) to call 'the father of modern ethology'.

Lorenz, who already at that stage had an unrivalled first-hand knowledge of animal behaviour, was the first to try to assess what existing science could contribute to our understanding of the normal, natural behaviour of animals under undisturbed conditions. After the publication of two penetrating studies of the social behaviour of Jackdaws (1927, 1931), and a more general study of social behaviour (1935—a paper that unfortunately reached the English-speaking reader in much too condensed, truncated form), he published in 1937 a critical assessment of the views of Spencer, Morgan, McDougall, and Ziegler, in which he showed that none of these could be reconciled with the new facts he presented. He then outlined what amounts to a modern version of Darwin's approach. In it he stressed the particulate nature of many behaviour patterns, and emphasized especially those relatively simple components that often appear at the end of a variable sequence of movements. During these introductory movements the animal seems to aim at finding the conditions in which these end acts, or consummatory acts (which he called 'instinctive acts') could be performed. In his largely inductive procedure he attached much weight to the following phenomena: the occasional 'misfiring' of instinctive acts under grossly abnormal circumstances; the phenomenon of 'explosive' or 'vacuum' behaviour shown when an animal has for a long time been denied the opportunity to perform an instinctive act; the evolutionary conservatism of instinctive acts, which makes them to reliable species and group characteristics comparable to structural characteristics; the internally controlled maturation of many behaviour patterns in the course of the development of the individual ('innate' behaviour); a learning process that he called 'imprinting' and compared with the process of inductive determination in the development of organs; and finally the mosaic-like integration of 'innate' and learnt behaviour components found in the adult animal as the result of unequal modifiability of these different components. Much of this was still highly tentative, and later research has led to certain modifications, elaborations, and corrections; yet this work was like a breath of fresh air and opened up vistas of unexplored territory. It is often forgotten (due in part to a certain assertativeness of Lorenz's style) that he explicitly wrote (p. 330): 'I do not see any danger in formulating my (perhaps rather heretic) views in this provocative, extreme way, as long as we are aware that they are working hypotheses that we should be prepared to modify if new facts force us to do so. One thing, however, I hope and believe to have shown convincingly: that the investigation of instinctive behaviour is not a subject for grand metaphysical ('geisteswissenschaftliche') speculation, but, at least for the time being, a task to be pursued by concrete experimental analysis.' (Free translation by me, N.T.)

It is interesting to see, in retrospect, how Lorenz's work gradually extended its influence, first into biology, then into psychology and psychiatry, and how, reciprocally, ethology responded to influences from outside. My sketch of this further development will naturally be a personal, and perhaps coloured, account.

Lorenz's new approach appealed at once to a small number of zoologists, namely those who, like myself, had already chosen to observe animal behaviour, and had been deeply disappointed when they looked for guidance at the psychological and the physiological literature. These observers too had been struck by the relative lack of modifiability of many behaviour patterns and their consequent conservative behaviour in phylogeny; by the apparent 'spontaneity' of much behaviour, which at that time seemed incompatible with reflex-theories; and by the fact that many behaviour patterns could often be seen to contribute decisively to survival, and to success in reproduction.

Lorenz's approach also appealed by its comprehensive nature—it was clearly a broad-fronted attack. Many more animal types were considered than before; ethologists aimed at describing and analysing all complex behaviour shown by animals in their natural surroundings; finally they were interested in function and evolution of behaviour as well as in its causation.

Before his work became more widely known, Lorenz elaborated (1939) two aspects of his theory that had not been given special attention in his first comprehensive theoretical paper quoted above. Both are important, because they had a bearing on the physiology of behaviour and led later to various contacts with neurophysiological work. Many non-conditioned behaviour components have been shown to be responsible to 'sign stimuli'—relatively minor parts of the total input with which the animal's sense organs could provide it. The organization of such behaviour patterns was obviously such that, under natural conditions, the sign stimulus was sufficiently characteristic of the natural objects to which the animal had to respond, to guarantee that the fitting response would occur. In some abnormal environments (viz. those that present the sign stimulus in a different context) this dependence on sign stimuli can lead to responses to the wrong object. This selective 'filtering' of sensory input, and therefore the effective sign stimuli, varies with the response shown, and often with the internal state of the animal. This type of evidence soon provided a link with a problem recognized as important by physiologists and cyberneticists: the centrifugal control (by 'gating') of the admission of input.

The second point of later contact concerned the internal, particularly central nervous control of behaviour. Independently, another individual 'heretic', the physiologist von Holst (e.g. 1939), had been demonstrating the existence of what he called spontaneous and automatic rhythms in central nervous activity underlying relatively complex behaviour patterns such as locomotion—movements not very much less complex than those that Lorenz had called instinctive acts. Von Holst's brilliant work in this field, which relied on the application of unconventional methods, culminated in his studies of 'relative coordination' of central nervous rhythms, and did much to arouse interest in the physiologically oriented aspect of Lorenz's work.

All the papers reporting on these new developments were written in German and published in German journals, and it was not astonishing that their influence remained at first confined to the German-reading parts of western Europe, particularly Germany and Holland.

II

The further development of ethology has been characterized by Koehler, who repeatedly pointed out that, like many new ideas, it went through three stages. At first, the new approach was largely ignored. Then, as its bearing upon wider issues began to be recognized, it met with vigorous criticism. This criticism led to further contact, and, through this, to a period of interchange of views and facts that marked the beginning of cooperation. This stage is often succeeded by one in which some notions are recognized as so self-evident that we tend to forget that they have not always been with us. In some fields of ethology we are entering this latter phase now.

The criticism was to be expected, because the notions developed by Lorenz and von Holst were in more than one respect contrary to prevailing thought, and so were bound to be unpopular. Von Holst's work was disturbing to those who were accustomed to think in terms of the reflex movement as the basic component of animal movement, as at that time most neurophysiologists were. Lorenz's notion of 'innate' behaviour clashed with the views of most American psychologists, who considered learning as the most important process controlling the development of behaviour of the individual—an over-emphasis that had undoubtedly much to do with the fact that animal psychology had derived its problems from human psychology.

Both Lorenz and von Holst expressed their views in an assertative, almost provocative way, reacting as they were against what they felt were strongly entrenched but biased positions. This made the ethological assault (and the counter-attack that it soon elicited) into an invigorating phase in scientific growth, characterized by a battle of wits fought with great intensity, from which, however, no side emerged loser or winner; all three disciplines profited by the confrontation.

The counter-attack opening the second phase came, after some skirmishes, when Lehrman published his '*Critique of the theories of Konrad Lorenz*' (1953), in which he made himself the spokesman of many of his colleagues among American psychologists.

Lehrman's criticism was in my opinion most incisive in questions of behaviour development. He argued that it is heuristically unhelpful to classify behaviour patterns, or parts of them, as either 'innate' or 'learnt'. Instead, the developmental processes ought to be analysed, even of those behaviours classified by ethologists as 'innate'. While such behaviour patterns admittedly developed without the aid of certain specified outside influences (for example when they were performed correctly before they could have been practised, or in response to stimuli which the animal had never met before), they might still be dependent on certain interactions with the environment, some of them different from those conventionally lumped under the name 'learning'. Lehrman claimed that the urge to disentangle such developmental processes is, so to speak, lulled to sleep by the rigid dichotomous classification of two behaviour types. Although he almost spoiled a good case by over-emphasizing the part played by early experience and by under-representing ethological work on behaviour ontogeny, later research has justified a great deal of what he said.

With regard to the physiological analysis of behaviour, Lehrman argued that ethologists often confused similarity in achievement of unknown mechanisms with similarity in organization of their machinery. While I think that he underrated the sophistication of the biologist precisely on this issue, his uneasiness was not entirely unfounded.

A third point of attack, related to both other points, was aimed at the optimism with which ethologists tended to assume similarity of mechanisms, both ontogenetic and physiological, between animal and human behaviour. As I hope to show later, modern ethologists, while still believing that Man and animals are more similar in their organization than many psychologists are prepared to admit, acknowledge perhaps even more readily than many psychologists the extent of our ignorance with respect to human behaviour, and in addition have shifted the emphasis of their argument to a point of methodology: ethologists now claim that they have developed *methods* which could fruitfully, and in fact should, be applied to human behaviour.

Finally, Lehrman showed the weakness of the evidence on which ethologists based their conclusion that autonomous processes in the central nervous system play a prominent part in the causation of behaviour. He made clear that they ignored the possible influence of one or more peripheral, in part sensory influences (for example proprioceptors, i.e. sense organs that report on the conditions and events inside the body rather than on the environment), which studies of intact animals could not properly elucidate.

Contact with physiologists developed a little later. The main reason for this seems to me to be the fact that both subject-matter and methods of the two sciences were originally too different. Neurophysiologists were at that time very much occupied with the analysis of processes at the level of the neurone or relatively simple neurone systems, and neither Sherrington's *Integrative action of the nervous system* nor Pavlov's work could be directly applied to complex behaviour. Von Holst, himself a pupil of the physiologist, Bethe, appealed at that time more to ethologists than to physiologists. Closely related to this difference in subject-matter was the difference in method. Physiologists were experimenters, and as such relied almost exclusively on the study of consequences of interference; as von Holst used to remark, this contains a methodological bias because by its nature the method deflected attention from what animals do when not interfered with, and thus from 'spontaneous' behaviour. Ethologists, faced with a paucity of purely descriptive data, tended to devote a large proportion of their time to observation, without as yet following up the resulting hypotheses with cogent experimentation. The gap, due to these differences in subject-matter, in method, and consequently in concepts and terminology, was wider than even those realized who saw the basic similarities in approach and believed in the ultimate fusion of all types of 'physiology of movement'. The early 'physiologizing' by ethologists was undoubtedly naïve; for instance the relevance of 'rest output' of sense organs to problems of 'spontaneity', and that of proprioceptive input to problems of drive reduction was not grasped, nor were we in general aware of the important concept of feedback. Reciprocally, it did not help our understanding of behaviour mechanisms when, for instance, Pavlov (1926) coined concepts such as ‘the curiosity reflex’ or when later Eccles (1953), on the basis of neurone studies, elaborated on the ‘mind’.

It was not astonishing therefore that contact between physiologists and ethologists developed relatively late, and that, when neurophysiologists began to take notice, they did so in a very critical state of mind.

The sharp distinction made here between psychology and physiology is of course highly schematic, particularly since American psychology had, by the time it came in contact with ethology, already developed physiological approaches and techniques. Yet the distinction is roughly valid because the centre of interest of psychologists was behaviour and that of neurophysiologists the nervous system.

Thus lack of experimental sophistication of ethologists, their concern with phenomena that were outside the sphere of interest of either psychology or physiology, and their interest in problems of survival value and evolution made it almost inevitable that their work elicited criticism from their fellow scientists; but it was equally inevitable that much of the criticism was based on misunderstanding. Under these circumstances, which could easily have led to a hardening of attitudes and mutual rejection of views and facts, it has been encouraging to see how soon the phase of *voicing* criticism moved into one of *listening* to criticism and, through this, to both a broadening of approach in all sciences concerned, and to convergence and fusion.

III

The post-war period has seen a rapid development of all the sciences concerned with animal behaviour. This development was undoubtedly in part due to trends in each of these sciences themselves. Psychologists were independently extending their interest to more types of animals, to more problems, and to the development of new techniques. Neurophysiologists began to extend their studies to higher levels of integration. Ethologists developed an increased tendency to experiment. It is worth stressing that this too was to a large extent due to Lorenz’s work. As one who has been closely associated with Lorenz since the mid-thirties, and who is experimentally inclined, I can testify that, although Lorenz is himself not an experimenter, his way of formulating his hypotheses has inspired and even dominated much experimental work, and is doing so even now, though often so indirectly that the younger generation does not always realize this. I think I can characterize his influence best by saying that, while the experimental method would have been applied to behaviour problems even if he had not provided his hypotheses, the approach would have been far less system-oriented and more atomistic in character. There is also no doubt at all that through his penetrating comparative studies he gave great impetus to research into the evolution of behaviour, including the effects of natural selection—a field still practically ignored by neurophysiologists, comparative psychologists, and human psychologists.

But a striking aspect of the post-war development has been the start of interdisciplinary contact: ethologists, while still continuing to call attention to what they are convinced is new in their work, are absorbing a great deal from sister sciences. Perhaps the most obvious development is the growing interest in studies of causation, which naturally moved in the direction of behaviour physiology. But ethology has retained a breadth of interest that is less pronounced in the other disciplines, and even in other biological fields. I believe that non-biologists, and even a number of my fellow biologists, do not always realize the full width of the major problems that biology, in all its ramifications, covers. With a slight modification of Huxley's reference to 'the three major problems in biology', viz. causation, function, and evolution, I have always found it helpful to think of biology as concerned, in 'commonsensical' terms, with two problems: that of causation and that of function in the sense of survival value. By this I mean that, starting from observable life processes, we ask 'what makes this happen?' and 'how do the effects of what happens influence survival (including reproduction)?' The first question can be roughly divided into three separate questions, differing in the time scale involved. The causation of short-term, cyclical events in an animal's life (such as periodic feeding, or the alternation between reproductive and non-reproductive seasons) is studied with a view to analysing the causation, the 'machinery' of such cyclical behaviour. Over a longer stretch of time, the whole life of an individual is considered to be one long cycle, and the changes in behaviour machinery in the course of the cycle, the development of the individual, has to be studied. Finally, as the generations follow each other, evolution takes place, in the course of which the ontogeny changes in successive generations. In spite of a certain degree of overlap, causation studies thus refer to three rather different problems. The study of the two former problems differs from the third one in that both are concerned with repeatable phenomena, which therefore can be submitted to experimentation with controls. But evolution is one single cycle, in which each step has been unique, and its study is therefore ultimately dependent on indirect methods, on reconstruction of what has happened. This is true even though, acting on the 'principle of actuality', (or constancy of the basic mechanisms throughout the past) one can study single components of the evolutionary process experimentally.

The problem of the functions served by life processes is in a sense of a different order. Organisms, 'building up negative entropy', seem to defy physical laws; they maintain themselves in what seems at first glance an unstable state. The fact that animals survive, reproduce, and evolve in the face of many opposing pressures makes us ask how they manage to do so. If survival of these systems—organisms or communities—were the observable from which we start, biology would be concerned with causation only, but since our observables are life processes, we have to follow up the cause–effect relationships in both directions; looking back in time when studying causation, forward when we study effects, in order to understand how these effects contribute to survival. Because, since Darwin, the organization of animals is assumed to be the result of natural selection, which favours the more successful forms, the investigation of the way in which life processes contribute to survival forms part of evolutionary study as well, and biology is therefore concerned with effects of life processes for two reasons. Of course, whether one traces causes or follows effects, the method employed is the same: the unraveling of cause–effect relationships, ultimately by experiment.

In a science such as ethology, which is characterized by an observable phenomenon rather than by, say, an animal type (e.g. ornithology), a problem (physiology), or a method (experimental embryology), an interest in all these problems is natural, and in this broad biological interest ethology still compares favourably with the bulk of physiological and psychological work.

In studies of the *causation* of short-term cycles (behaviour physiology) Lorenz's original views have been partly elaborated, partly corrected, partly supplemented by notions that allowed new questions to be asked. Lorenz described how functionally unitary behaviour sequences (such as the different activities that together are called feeding) often consist of an introductory phase, which (following Craig 1918) he called 'appetitive behaviour' and is characterized by a high degree of 'spontaneity' (i.e. internal initiation), by irregularity or variability of motor patterns employed, and by the fact that it is continued until the animal meets with a situation that elicits the next link in the behaviour chain. Thus a hungry animal sets out 'in search of' food, and the switch to hunting, eating, etc. occurs when food is actually perceived. He further called attention to the fact (likewise pointed out by Craig) that such sequences often end in relatively stereotyped end acts, or consummatory acts ('instinctive 'acts), and that the readiness to repeat or continue the same type of behaviour drops more or less suddenly with the actual performance of the end act. He thus stressed the facts that behaviour is in part internally motivated; that activities that belong together functionally are also connected causally in such a way that they occur together in certain sequences; and that the entire complex, so to speak, loses its internal motivation after performance of the end act(s). This led him to introduce the concept 'action specific energy', a more or less metaphorical term for largely unknown agents that influence specific, functionally distinct behaviour systems rather than general activity. His studies of fluctuations of the readiness to perform a certain behaviour showed that if an animal is denied the opportunity to show such behaviour, it becomes progressively more ready to respond to incomplete stimulation, and when stimulated it shows increasingly complete behaviour. In extreme cases, behaviour that is normally dependent on elicitation by stimuli, will appear 'spontaneously' ('vacuum activities'). Lorenz went one step further and identified 'action specific energy' with one particular type of internal control, viz. control by processes in the central nervous system itself, which he thought might be akin to the 'automatisms' in von Holst's sense. Rather ignoring the possibility that the drive-reducing effect of performance of the end act might occur through the medium of, for instance, proprioceptive stimuli, or even outside stimuli, he ascribed drive-reduction to direct depletion of 'action specific energy'. The model he proposed as an analogy did much to establish an image of his theory that was narrower and more rigid than the theory itself. It consisted of a tank slowly being filled with water ('action specific energy'), which could be emptied through various outlets, each of which could be opened by a trigger device ('releasing stimuli') that allowed the water to run out ('consummatory act'). If no triggers were pulled, the water would overcome the resistance blocking an outlet ('vacuum activity').

Lorenz's emphasis on internal behaviour-specific determinants has undoubtedly been of great importance, even though his specific hypothesis was too narrow. For instance, new work on the 'rest output' of various sense organs—an output affecting the central nervous system even when no specific stimuli are administered—indicates that even specific behaviour patterns such as coordinated swimming may depend on such unspecific input. The exact source of input need not be important, but when it is completely absent swimming stops (e.g. Lissmann 1946). Closely related to the notion of spontaneous central nervous rhythms is the idea that the well-known influence of certain hormones on behaviour is a direct one. Later analyses have shown that this is not invariably true: reproductive hormones may cause growth processes, for instance of peripheral sense organs, which as a consequence become capable of stimulating the nervous system. However, direct effects of hormones on the central nervous system have been demonstrated, and the main area of impact has even been identified as the posterior hypothalamus (Harris, Michael, and Scott 1958). At the same time, neurophysiologists have demonstrated in an increasing number of cases that central nervous tissue does much more than meekly wait to be prodded into action. The old notion is being replaced by one envisaging continuously active populations of nerve cells that are usually inhibited, but when disinhibited by other centres or by sensory stimuli give rise to coordinated behaviour (see, for example, Roeder 1962). This level of activity depends on a variety of (partly extraneural) agents, and the 'endogenous' contributions of the central nervous system are now being studied in terms of the relationships, often the discrepancies between the inputs and the outputs —in short on the *processing* done by the central nervous system.

Similarly, Lorenz's emphasis on the drive-reducing effects of performance has stimulated research, and through this has led to a more comprehensive view of the problem than he presented at the time. In particular, his 'psychohydraulics' model differed in one important respect from the biological systems on which it was based: its self-regulatory properties were not sufficiently emphasized. Under the influence of cybernetics (which reached ethology at second hand, via engineering science and physiology) it was recognized that, when performance of the end act 'as such' reduced drive, performance must provide some kind of negative feedback. Once the search for feedbacks began to be applied to major behaviour systems, it was found that feedback loops, involved in drive-reduction after performance of a consummatory act, could vary widely from one case to another. Thus satiation after feeding is at least in part under the control of negative feedback from one part or another of the intestinal tract (e.g. Berkun, Kessen, and Miller 1952; Deither and Bodenstein 1958); the reduction of sexual behaviour after sperm ejaculation was shown in a fish, the Stickleback, to be a consequence of stimuli from the eggs rather than from the loss of sperm (Sevenster-Bol 1962); a particularly extensive feedback loop since the eggs are laid by the female in response to behaviour of the male that precedes actual sperm ejaculation. In these cases, the animal responds to 're-afferent' stimuli, i.e. stimuli that report back on the effects of the activity performed. On the other hand, the autumn migration of Starlings was shown by Perdeck (1964) to end after the birds have flown a certain distance or a certain time, not when they have reached a certain end situation. Blest (1957) analysed, in a moth, the reduction of the readiness to fly after having flown, and found, by elimination of all conceivable extra-neural feedbacks (including those from proprioceptors) that the feedback must arise within the central nervous system itself—an example even more fully in accordance with Lorenz's ideas than that studied by Perdeck.

The analysis of the organization of complex, functionally unitary behaviour patterns is likewise being pursued and leads to a more complex and varied image than that originally proposed by Lorenz. The basic fact that functionally-related behaviour components are appearing in certain orderly patterns has been established with much greater accuracy, one of the methods being applied being factor analysis (Wiepkema 1961). In general terms this type of integration of functionally related components into systems means that the components must be causally related *somehow*, but the organizational principles involved are much more varied than that suggested by concepts of unitary drives in the sense of superordinated motivational factors controlling directly all components—perhaps with relatively simple threshold differences, or different relations with qualitatively different stimuli for each component. The web-like character of the mechanisms underlying major behaviour systems is now being shown to be more complex; they include many feedback relationships and mutual (facilitating and inhibiting) interrelations at and between a variety of levels. The work on the organization of reproductive behaviour in birds (Lehrman 1961; Hinde 1965, 1966), for instance, has now moved far beyond the demonstration that reproductive hormones are required for their appearance. Hormones influence growth processes not only in other hormone-producing organs but also in sense organs and in effector organs, with numerous feedback links and interrelations between separate cause–effect chains. Conversely, the activity of the hormone-producing glands can be affected by external stimuli, part of which in turn arise from earlier hormone-controlled effects. Thus the endocrine cycle of female pigeons is affected by stimuli provided by a courting male. Nest building in Canaries, controlled by the female sex hormone oestrogen, leads to the completion of the nest cup; this not only stops this phase of building but also reduces the further secretion of oestrogen. In the meantime the broodpatch on the female's underside develops under control of oestrogen together with a secondary hormone, and as a consequence becomes more sensitive to touch stimuli. This makes it send back strong touch stimuli when the female sits in the cup, and this again makes the bird switch to softer material, i.e. feathers, with which the nest cup is then lined. The diagram by which Hinde (1966, p. 302) summarizes the relations so far known or suspected, in both a positive and negative sense, between the hormones, the behavioural and the physiological changes, and outside stimuli, shows a highly complicated web.

The influence of reproductive hormones can thus be exerted in indirect ways, for example through growth processes, but also directly, as hormone-implantations in the brain have been shown (Harris, Michael, and Scott 1958). Brain-stem stimulation experiments further show that electrical stimulation of very simple rhythms can produce behaviour patterns of great complexity, suggesting hierarchically organized executive systems (e.g. von Holst and von St. Paul 1963).

It is difficult to assess what contribution the concept of 'action specific energy' has made to these developments. It has certainly drawn attention to problems of the internal control of major behaviour patterns, and although it was clearly an exaggeration and simplification to ascribe so much to more or less autonomous central nervous mechanisms, both the 'spontaneous' and the integrative activities of the central nervous system are clearly much more important than was recognized at the time. The concept of unitary drives has further rendered excellent services in the first steps of the analysis of 'conflict behaviour', i.e. behaviour such as 'threat' and 'courtship' resulting from the simultaneous elicitation of two or more major behavioural systems (e.g. Tinbergen 1964). For purposes of more detailed analysis, however, knowledge of the organization of functionally unitary major systems becomes increasingly important.

Increased attention is also being given to the problem of internal control of sensory input. It has long been recognized that the responsiveness of animals to external stimuli varies, and must be under internal control, but analysis of the processes involved has not started in earnest until recently. Physiologists now begin to trace the messages going out from the centres towards the sensory periphery which can 'set' the sensitivity of sense organs and can even 'filter' part of the sensory input. Such centrifugal control of sensory input by inhibition may well be effected in a variety of ways and at different levels; in the ear of the cat it seems to be the nervous cells in immediate contact with the sensory cells which, by the outgoing nerve known as 'Rasmussen's bundle', can be prevented from passing on auditory input (Desmedt and Monaco 1961).

Equally interesting is the widening approach to 're-afference': the processing of stimuli from the sense organs which are received as a consequence of an activity of the animal, and by which it judges what effect the movement has had. Of particular interest are those studies in which the interaction between centrifugal setting and re-afference is investigated. The re-afferent stimuli have much wider effects than that of reducing 'drive'; they are, inside the animal, compared with something like a 'template', a representation of what these stimuli 'ought to be'. It is the *relation* between template and actual re-afference that dictates which behaviour shall be shown subsequently. Von Holst and Mittelstaedt (1950) have demonstrated by a simple experiment that such a template can represent quite complex features, and also that it can vary with the internal state of the animal. This experiment is worth being described in some detail. When an animal moves in relation to the visual environment (either because it moves itself or when the surroundings move) the image of the environment moves over the retina of the eye. When the animal is stationary and receives such a stimulus, it responds by correction movements (the 'optomotor response'), which restore its original position in relation to the environment. However, when in a stationary environment the animal moves on its own accord, the retinal image moves as before, and yet the animal does not then respond by moving back. Von Holst and Mittelstaedt showed in a hover fly that in the latter case the optomotor response is not simply inhibited, but that in both cases the animal receives re-afferent messages from the eyes. Correction movements follow only when these 'checking-up' stimuli do not tally with what could be expected: when the animal is stationary, retinal movement is not expected and must mean movement of the environment, but in the moving animal the retinal image *should* move, and the centres take this into account. In other words, at each moment these receiving centres are informed about the 'Sollwert', i.e. the value that the re-afferent stimulus 'ought to' have—in functional terms: its 'neutral value', the template into which this re-afference should fit; actual re-afference and Sollwert are *compared*, and discrepancies between the two lead to correction movements. The experiment shows that the value of such a template is 'set' to correspond with movements that the animal is performing. Similar processes can easily be shown to occur in our own visual system, and they are no doubt of very general occurrence.

These and many other investigations are exploring a field that lies, so to speak, half-way between ethology and neurophysiology. Other such researches concern the integration of separate sensory data into unit-messages of a higher order, such as is the case in vision of form (e.g. Sutherland 1962; Waterman, Wiersma, and Bush 1964) and of movement (e.g. Hassenstein 1961). The result of this widening of the front of attack is that the distinctness between some fields of ethological, psychological, and physiological research is beginning to fade. This was to be expected, because these phenomena of different levels of integration are being studied with essentially the same method—although of course the special concepts, techniques, and terminologies remain different, adapted as they have to be to the level studied.

Studies of behaviour *ontogeny* have likewise made considerable progress, but there is still profound disagreement between American psychologists and (particularly German) ethologists on the methods of approach, and even on the aims of research. This is clearly demonstrated in two recent treatises by prominent representatives of these two groups, Lorenz (1965) and Schneirla (1966). English ethologists occupy on the whole a middle position.

The dispute started when Lorenz (1931, 1935) described many examples in which animals show species-specific, complex behaviour patterns (often in response to specific stimulus situations) in circumstances that had precluded the opportunity either for learning by practice or example, or for conditioning to the effective stimuli. He criticized the tendency among American psychologists to assume that learning processes, and other types of interaction with the environment during growth, were the main or even exclusive determinants of ontogeny. By classifying behaviour patterns in two ontogenetic types, innate and acquired behaviour, Lorenz stressed the importance of internally controlled development.

In subsequent work, the two parties applied different methods. On the whole, the ethologists proceeded by demonstrating that certain learning processes, at first glance likely to take part in the moulding of such responses, were actually not involved in a number of examples. Thus Grohmann (1939) showed that the incipient, gradually improving flight movements made by nestling doves were not influential in the development of the basic flying ability of fledglings; likewise many responses to intra-specific signals were shown to be unconditioned (summary, for example, in Tinbergen 1953); and many more similar examples are known. Psychologists naturally studied the effects of various manipulations of the environment in which animals grew up, and, equally naturally, tended to emphasize those cases in which their search was rewarded. Thus, undeniably, ethologists studied a broader spectrum of phenomena, whereas psychologists penetrated more deeply into problems of learning.

Closely related to this was a difference in aims, which has only gradually become clear. Already Lehrman (1953) argued that the analysis of the whole developmental process, from very early stages on, and with reference not only to learning but to other formative outside influences as well, should be the real aim. Lorenz made it clear recently (1965) that '*Not being experimental embryologists* but students of behaviour, we begin our query, not at the beginning of the growth, but at the beginning of the function of . . . innate mechanisms' (p. 43). (Italics mine, N.T.) With Lehrman and many English-speaking ethologists I believe that we *should* be concerned with the entire developmental process—that behaviour development will only be fully understood if we apply all the methods of experimental embryology to problems of behaviour development.

In view of these differences of attention, of aims, and of methods, no useful discussion is possible without a prior consideration of some matters of semantics.

Male three-spined Sticklebacks, raised in isolation from fellow members of the same species, show, when adult, normal fighting behaviour in selective responses to other males (e.g. Cullen 1961). This is, in Lorenzian usage, an 'innate response'. Strictly speaking, the term here means a response neither influenced by the example of an adult male (a possible 'teacher'), nor by practice, nor by conditioning. It means a *non-learnt* response, and this, as Beach has said long ago (1955) is a negative definition. To workers interested in the entire developmental process, the word 'innate' implies no interaction with the environment at all; not 'acquired' through such interaction, at any stage, at any level. Now it has been shown for instance in tadpoles that exposure to light at an early stage is necessary for the rods in the retina (one of the two types of light sensitive cells) to develop their proper function (Knoll 1956). While this cannot be called learning, it *is* an interaction with the environment, which *is* required for the proper development of this component of the visual system. In the stricter sense, therefore, none of the subsequent visual response of the tadpole, nor of the adult frog, can be said to be innate, even though many of these later responses may not require learning processes. As long as the earlier stages of development have not been investigated, one has not eliminated possible interactions with the environment. This kind of gap in the evidence seems to me to be at the root of the conviction of many psychologists and English-speaking ethologists that it is not helpful to use a rigid dichotomous classification of innate and acquired *behaviour*, unless one qualifies the word 'innate' rather drastically, as indeed Lorenz does.

In other contexts however the word 'innate' is not objected to. When two groups of animals grow up in the same environment (admittedly a requirement that is difficult to meet in practice) and develop different behaviour repertoires, these *differences* are in the strictest sense of the word innate; to be more precise, they are due to genetic differences between the groups, and the ultimate aim of the ontogenist is to analyse where in the developmental processes of both, and how, the different genetic instructions lead to different results. The same formulation can be applied to similarities between animals growing up in different environments.

The dichotomy is further (as Lorenz has pointed out in 1965) completely justified with respect to the *sources of 'programming'* of the animal. The information comes either from within the animal or from the outside world. Male ducks of many species have to become conditioned to the females of their own species at an early age ('imprinting'), but females of these species mate preferentially with males of their own species even if they have been raised with males of another species (Schutz 1965). Of course internal programming is in its turn the result of the trial-and-error interaction with the environment which directs evolution and which we call natural selection.

It is very remarkable indeed that in this confused situation the actual research done by ethologists and psychologists is so very similar. The information, referred to above, that tadpoles require light for their visual system to develop fully, comes from a zoologist; a great deal of work on the genetics of behaviour differences is done in American psychology laboratories. This must mean that much of the disagreement is due to differences in formulating aims and conclusions, and in differences of degree in the guiding interest, which expresses itself in the concentration of psychologists on environmental aspects and of ethologists on internal aspects.

The procedure now applied by many workers could be roughly characterized as follows. Descriptive studies of the behaviour of developing animals reveal how the behaviour 'machinery' changes in the course of time. By comparing the effect of different early environments on the development, certain environmental features are either demonstrated to be effective or shown to be uninfluential. The subsequent analysis of the processes involved takes different courses for environmental effects than for internal aspects of programming. I think it is often forgotten by champions of 'innate behaviour' that, in order to get away from demonstrations that certain environmental aspects are *not* influential, one has to analyse internal processes directly, and this can only be done by interfering with processes inside the animal. On the other hand, champions of 'acquired' behaviour often seem to forget that 'acquisition', including learning, is not, so to speak, creating something out of nothing; it is a process of *changing*, and often perfecting, through interaction with the environment, something less perfect that was already functioning before. Thus the social interaction between a mother cat and its kittens begins with certain unconditioned responses in both, which improve by a series of adaptive adjustments made as a consequence of mutual interaction (Schneirla, Rosenblatt, and Tobach 1963). Another example is the development of song in Chaffinches and other birds. Males raised without being able to hear the song of older, experienced males do not develop the normal song, but they do develop a simpler, warbling song (Thorpe 1961).

As has been already indicated in the example of the tadpoles, it is further of great importance to distinguish clearly between stages of development and between the levels of integration involved.

The relevance of such facts to our problem could perhaps best be made clear by the following analogy. If, in the course of constructing a piece of scientific apparatus, I ask a factory to cut a sheet of metal into the desired shape, which I subsequently bend before incorporating it into my apparatus, is this apparatus factory-made or home-made? With respect to the metal part itself, would it make sense to call *it* either factory-made or hand-made?

A few examples must suffice to show how a picture is gradually emerging that shows promise for our ultimate understanding of the developmental process; a picture much richer than a mere dichotomy into innate and acquired responses suggests. This is not to say that Lorenz's emphasis on the extent of internal programming has not been important—in fact I think that, particularly with the extension he has recently (1965) added by introducing the concept of the 'innate teaching mechanism' (see below), it will have very far-reaching consequences even for our understanding of human behaviour. But it is worth pointing out that we have so far not agreed on a sensible set of concepts and terms.

The considerable promise of direct attacks on internal processes has been demonstrated by the rightly famous experiments of Sperry and his collaborators (see, for example, Sperry 1959). In the normal development of a vertebrate the sensory nerves that report messages from the skin grow out from the spinal cord towards the skin areas they serve. When, at an early (yet not too early) stage of development of a tadpole a piece of prospective dorsal skin tissue is transplanted to the ventral side of the embryo, it develops into a patch of dark-coloured, i.e. dorsal skin, so that the adult frog has on its belly a piece of skin that has the characteristics of a piece of dorsal skin. Conversely, prospective ventral skin that is transplanted to the back (and in practice both manipulations are done at the same time by exchanging the two pieces of tissue) likewise retains its characteristics, and so appears later as a piece of light-coloured skin on the back. When these transplantations are carried out before the sensory nerves have grown out to the skin, the nerves that establish contact with the grafts are those that would normally have contacted the skin areas now occupied by the transplants. When now the adult frog is touched on one of the transplanted skin areas, it makes, as any normal frog does, scratching or wiping movements with a hind leg. But unlike normal frogs, such a frog wipes its *ventral* side when the piece of white skin on its *back* is touched, and it wipes its *back* when the piece of dark skin on its *belly* is touched. The conclusion is inescapable that the nature of the skin, even when it grew up in the wrong place, determined the later function of the sensory nerve and the reflex mechanism it serves. The stage in the developmental process in which the function of the nerve is determined is therefore under internal control with respect to the animal as a whole (although, since the skin exerts an influence on the nerve, we cannot call the process 'internal' with respect to the central nervous system). Sperry suggests that similar internal formative processes may well take place in a staggered series of events inside the central nervous system itself.

Since the development of young animals cannot be sharply distinguished from such processes as the seasonal development of the reproductive condition, the work of Hinde and Lehrman mentioned before, which is concerned with another set of internal determinants, the hormones, is relevant here too. Hinde's work in particular shows how fruitful it is for such studies to analyse the interaction between various parts of the body, the stage-by-stage nature of the entire process, and the numerous antagonistic, synergistic, and feedback relationships within the system, and between the system and the environment.

Of particular interest are analyses in which the non-learnt components of complex behaviour chains are further programmed and integrated by interaction with the environment. It has been found in several instances that animals raised without opportunity to learn a complex behaviour sequence do develop the elementary *components* of such an action chain, but have to have experience with the natural objects in order to mould them into an efficient skill. For instance, Squirrels crack hazel nuts very expertly, by first manipulating them into a position in which they can gnaw selectively along the pre-formed groove where the shell is thin; then quickly weakening this groove, and finally cracking the shell along this weakened line. Squirrels raised alone and without nuts develop manipulating, gnawing, and biting, but their sequences are irregular and, because they gnaw all over the nut's surface, it takes them a long time to weaken the shell so that it can be cracked. Very long practice with nuts is required to develop this high-level skill (Eibl-Eibesfeldt 1963).

Naturally, such analyses are only first steps, and while interference with processes inside the animal is required to further analyse the internal, 'maturational' aspects of development, further manipulation of the environment is used to determine which outside influences are effective and in what way they change the behaviour machinery. Thus it is important to know what exactly reinforces a response (for a recent summary of evidence see Hinde 1966).

Both psychologists and ethologists have followed up this problem, though each in their own way, and both come to the conclusion that even where learning and other interactions with the environment occur, this learning is selective; for instance learning by sheer 'contiguity' of a response and any resulting stimulation is no longer believed to be a generally valid principle. Ethological evidence is mounting which shows that different species, even in roughly comparable situations, learn different things, and that different responses of the same animal are changed by different aspects of the environment. For instance, recognition of (i.e. selective responsiveness to) the correct conspecific sex partner is not conditioned in many species, whereas in Schutz's drakes it is. However, as we have seen, the *females* of these same species of ducks respond selectively to males of their own species even if raised with a different species. Numerous other examples of such 'predispositions to learn' are known.

This selectiveness of learning is also shown in such studies as Thorpe's, whose Chaffinches do not acquire just any song they hear but such that conform in certain aspects to the song of the species. The concept of something like a pre-set 'expectancy', which makes the development of an animal at some stages particularly responsive to specific stimuli, is required, and is being introduced under a variety of terms, such as the 'innate teaching mechanism' (Lorenz 1965), or a 'template' (Thorpe 1963) against which the animal matches the incoming stimulation. However, the concept should certainly not be confined to *innate* teaching mechanisms—the template may itself be the result of previous learning processes. Thus further analyses of the way in which some birds acquire their song has shown that they may, by listening to experienced singers during an early critical phase, establish a template on the basis of this experience, to which they adjust their later song by comparing the sounds they make with it (Konishi 1965).

Thus the conceptualization in studies of behaviour development is evolving along lines strictly parallel to those in short-term causation studies. In addition, a process of interdisciplinary fusion is discernable in both fields. Lorenz began, in reacting against an over-emphasis of the part played by external control, *by classifying*, stressing the fact that part of the total programming of an animal's behaviour machinery is done within the animal itself. By applying the term 'innate' to a class of behaviour patterns, however, he made it difficult, I believe, to move away from the preparatory, and at the time useful, classification of behaviours, and towards a programme of detailed analysis of the developmental processes. Yet it has been of great importance to realize that, by his own work and that of his pupils, he did call attention to the variety of patterns of interplay between internal and environmental programming. The analysis of the actual processes involved in internal programming has hardly begun, and so far many conclusions about such internal processes can only be supported by negative, eliminative evidence. But again, this evidence is not to be discarded lightly. For instance, while one could argue, as Lehrman (following Kuo 1932) did, that the pecking movements of newly-hatched chicks could have been influenced by the head having been moved passively by the heartbeat while the chick was still in the egg, how could one possibly imagine a duckling having acquired, by any interaction with the environment, the peculiar movements of the wing that ducklings may make while fighting, plus the details of its orientation, which, when the behaviour first appears, is already 'calculated' to the size that the wing will not acquire until later? Facts such as those reported by Thomas and Schaller (1954), that kittens raised without a mother and in complete darkness show the complicated 'hunting plays' the first time they are shown a dummy mouse; that Sticklebacks reared without conspecifics show normal and selective fighting and courtship behaviour (Cullen 1961); that young Blackbirds exclusively fed by white objects gape selectively at black dummies (Tinbergen and Kuenen 1939); that male Chiffchaffs and Grasshopper Warblers raised in isolation produce (unlike Chaffinches) their normal, admittedly not very complex song (Heinroth and Heinroth 1928)—such facts all point to important internal contributions to behaviour programming. On the other hand, the preoccupation (which I sense in workers such as Schneirla even now) with the possibility that the search for interactions with the environment at early stages will always show such interactions, prevents such workers from appreciating that not all the programming can possibly come from outside. Thus both the eliminative procedure and the restriction to manipulation of the environment fail to demonstrate, let alone analyse, the internal processes, as research of the type done by Sperry does, admittedly so far only for relatively simple levels. The true story of behaviour ontogeny will only be discovered by studying external and internal events and their interplay.

Of course these developments are of the greatest importance for an understanding of the ontogeny of our own behaviour. The experiments that are required can, for ethical reasons, never be done with human beings; deliberate interference with a child's development of the types and to the extent required is morally unacceptable. Clinical evidence is the best we can expect (see, for example, Bowlby 1960); it should be collected on the largest possible scale and checked against the results of animal experiments. Ethologists claim that it is by no means proved, and is in in fact highly unlikely, that manipulation of the environment (education in the widest sense) can mould Man's behaviour beyond the boundaries of innately determined ranges, although the extent of these ranges are hardly known. At the moment, it is neither scientific to claim actual knowledge of our innate behavioural equipment nor that we are infinitely mouldable—that, say, our aggression can be eliminated entirely by educational measures; such questions have to be considered undecided until they are properly investigated. The importance of modern work on animals lies at least partly in the fact that it is beginning to give us the tools required for such studies.

Unlike behaviour physiology and behaviour development, the study of behaviour *evolution* is still entirely in the hands of zoologists, neither psychologists nor physiologists having so far shown much interest in the field. Here ethology is widening its approach through increasing contact with ecology, genetics, and evolutionary studies in general—sciences developed by other biologists. The classical methods of genetics are now beginning to be applied to behaviour; correlations are being established between genetic, structural, and behavioural differences of different strains, populations, subspecies, and species; the inheritance of behavioural traits is followed through one, two, or more generations following cross-breeding, etc. The field, however, is still in its infancy, even though a large number of such correlations have been established at various levels of behavioural complexity. The gap between knowledge on the level of the genetic 'blueprint' and that of behaviour is still very large, and will only be filled by laborious analysis of behaviour machinery and its ontogeny.

At the same time, the part played by natural selection in moulding behaviour, as well as that of behaviour in creating new opportunities for natural selection, are being explored.

With regard to the first problem, several approaches are being adopted.

Natural selection is being applied artificially and its consequences studied. The grandest experiment done so far was the domestication of various animals, of which, however, the scientific documentation has been extremely sketchy. But refined measuring techniques and the use of fast-breeding animals such as the fruitfly *Drosophila* are now enabling us to apply selection pressures over a number of generations and to study its results within a time span of the order of a few years. The evolution of sexual isolation (i.e. lack of cross-breeding) between related strains—a phenomenon that is at the root of evolutionary divergence and is often a matter of mating behaviour—is an example. Species often split up into populations of slightly different constitution when they increase their geographical area. In the different corners of such an area populations go their own evolutionary way, partly because they are derived from slightly different, not quite representative samples of the heterogeneous original population, partly because, living in different environments, natural selection imposes additional differences. When such populations, in the course of their subsequent movements, come into contact again, their representatives may still cross-breed. But due to their different genetic makeup, the hybrids are usually less viable than both parent strains. In other words, natural selection discriminates from this moment against those members of the parent strains that cross-breed, by penalizing their offspring more than pure offspring of either strain. Experiments are being done to test the hypothesis (derived from field observations) that such anti-hybrid selection favours, in both strains, the further development of mating preferentially within the strain. The results so far obtained are according to this hypothesis, that is, sexual isolation increased. Moreover this was shown to be due to certain evolutionary changes in mating behaviour (Crossley 1963).

There is also an increasing number of studies (see, for example, Cain 1964, Tinbergen 1965 *b*) that are testing whether the behavioural characteristics of species are really such that they make each species singularly well-adapted to the ecological 'niche' in which it is living, and if so, which pressures of the environment it meets, how these pressures exert their influence, and finally how the animal deals with them. These studies are very similar to studies of survival value and the adaptedness of behaviour pursued in their own right, as mentioned above; the difference lies in their ultimate aim. Taxonomic and geographical studies can reveal the way in which populations have in the past expanded, split up, often subsequently overlapped, etc. When such studies are combined with ecological work, in particular investigations that test whether invasion of new habitats, of life in changing habitats, etc. produce changes in structure and behaviour of populations, one gets an idea of the extent to which natural selection must have been effective.

At the same time experimental studies are testing whether the behaviour differences between different species are actually adaptive, as they should be if selection has moulded them. This is done by comparing the success of a roughly normal population of a species with a population which in one of its characteristics deviates from the norm in the direction of another species. For instance, the several species of gulls differ in that while most species remove the empty egg-shell after each chick has hatched some species leave the egg-shell in the nest. It has been shown that egg-shell removal is a corollary of the camouflage of the brood: eggs and chicks are protected by their colour against predators that hunt by sight, while the broken egg-shell, which shows the white inside and edges, attracts predators, which then find and eat the brood (Tinbergen *et al.* 1962). The species that do not remove the egg-shell seem to have no need for it. For instance, one such species, the Kittiwake, nests on narrow ledges on vertical cliffs, where its broods are practically out of reach of predators.

The studies so far done along these lines begin to show how beautifully and intricately the behaviour of each species is adapted to its needs—within, of course, the limits of the overall capacities of animals at each evolutionary level. There are several circumstances that explain why our knowledge in this field is still meagre, and why the power of natural selection in moulding behaviour is still much underrated. First, few biologists are engaged in this type of research; the successes of the physical sciences having drawn much of the available talent to physiology, biochemistry, and biophysics. Second, our knowledge of the many pressures that the natural environment exerts on each species is still extremely poor. Third, the analytical method forces us to select for study one behaviour characteristic and one environmental pressure at a time, and one of the very natural reactions after one such study is to think that an animal could perform a particular task much better than it does. However, where this question has been studied in a broader context it has always become clear that there is, within the animal, competition between various activities, and that different pressures require different and not always compatible ways of meeting them. To mention a simple example: many young birds are camouflaged as a protection against certain predators. Camouflaged colour patterns are effective only when the animal is motionless. However these animals have to eat, and this requires motion. Thus their behaviour has to be a compromise; for instance, they will feed normally, and will 'freeze' when the parent spots a predator and calls the alarm. While they could feed more efficiently if they never had to freeze, and would be better protected against predators if they never had to move, they can do neither, and selection, rewarding overall success rather than any isolated characteristic, has produced compromises. A full understanding of the total problem of adaptedness therefore requires a synthesis of data concerning single behavioural traits and single environmental pressures.

The ways in which behaviour *creates* new opportunities for natural selection are also being studied. Perhaps the most basic behavioural contribution to evolution is the fact that every animal population continuously explores new habitats. Floating larvae of marine animals, and 'dispersal stages' of many land animals (for example spiders sailing on the wind with the aid of gossamer threads) are year after year carried in enormous masses outside their optimal habitat; young individuals of territorial animals are often repelled from already occupied habitats and forced to try to settle elsewhere. The majority of such 'pioneers' dies, but changing conditions in such newly explored habitats, or changes within the animals themselves, lead again and again to the invasion of previously unoccupied areas. Whenever this happens, the new environment inexorably enforces further evolution. The most famous example of this is provided by the South American finch-like birds that have invaded the Galapagos Islands, where they have developed into a group of widely divergent species (Lack 1947).

Very spectacular examples of evolutionary change due to behaviour are found wherever the environmental pressure acting on an animal is exerted by another animal, such as in prey–predator relationships, or in social relations within the species, for example relations between sex partners or between parents and young. In relationships within a species, where success is often dependent on a fitting signalling system, it is usually difficult to say whether it is the signal that has caused the recipient to evolve a specific sensitivity to it or whether specific sensitivity of the recipient has enforced the development of a signal such as a movement, a sound, a scent gland, or a brightly coloured structure. But in some cases of intraspecific signalling this can be established. Wickler (1962) has described a species of mouth-breeding fish, in which the male has, on its anal fin, a series of colour spots that are remarkably accurate two-dimensional 'pictures' of the eggs of the species. During mating, the female takes the eggs into her mouth immediately after laying, but for the eggs to be properly fertilized she has to snap up the male's sperm as well. She is made to do this by the male's displaying the egg-'lures'; in her attempts to snap these up she takes up the male's sperm. Clearly the evolution of the signal is an adaptation to the female's visual response to eggs, which in its turn is an adaptation to the eggs. Numerous examples are known in which behaviour of a predator forces the prey species to evolve defences, a simple case being the general motionlessness of camouflaged animals (movement being a powerful stimulus that would destroy the camouflage effect); more complex are the 'distraction displays' by which many ground-breeding birds lure predators away from their brood. In the struggle between predator and prey the prey species can also direct the evolution of the predator; thus various predators have developed lures that mimic the food of their prey species.

IV

In this admittedly sketchy outline of the development of ethology to date I may well have misjudged the relative importance of contributions made by this young science in comparison with those of sister disciplines. The exchange between these various sciences, in published work and, more importantly, in many personal contacts, has been so intensive and in addition so interwoven that it is impossible for any one person to keep a good or even an objective record. Moreover, many changes of viewpoint and of approach have occurred at an intuitive, unconscious level and have not been made explicit until at a late stage—they were expressions of general trends that were 'in the air'. It seems justified, however, to say that the behavioural sciences have in these sixty years moved gradually towards increased affinity with other natural sciences, and that ethology has contributed substantially to this development. In spite of its initial bold simplifications, its extreme positions in reaction against prevailing views, and its various other shortcomings, and in spite of the fact that the other behavioural sciences had moved independently towards both more scientific and broader methods of approach, I believe that it is fair to say that ethology has had a healthy, invigorating effect, not only on this process but also on hastening the fusion of many separate disciplines into one comprehensive main stream of truly biological research.

The future consequences of this development will undoubtedly be of great importance. The methods and concepts of this emerging biological science of behaviour are proving themselves so successful in animal studies that they will have to be applied to human behaviour as well; in fact such studies are already starting. Few will deny that there are many disturbing signs of malfunctioning of our own behaviour, particularly of our social behaviour. It is a task of the greatest urgency to try to find out how this came about. The most likely hypothesis is that the culturally determined changes in our environment (particularly our social environment) have outpaced adjustments in our behaviour; that genetic evolution is much too slow to achieve such adjustment; and that our individual behaviour is not sufficiently modifiable because we too are genetically restricted—a consequence of our still being adapted to the ancestral environment. Animal behaviourists agree that the conception of Man as an infinitely adjustable species, of which each individual can in its lifetime be modified behaviourally by educational measures to any desired extent, is not necessarily correct; in fact it is highly unlikely to be true. And even if it were largely true, it would still be abundantly clear that, in our ignorance of the natural control of behaviour ontogeny, we have not yet found the best ways of guiding the development of young individuals.

This consideration is relevant to a sketch of the growth of our science because further development may well be hampered by a kind of social negative feedback, a 'backlash'. As the scientific understanding, and through it the control of our own behaviour, will be seen as even a remote possibility, the resistance against this type of self-analysis may well increase. Whether such resistance can itself be eliminated by education is perhaps an even more basic problem, but even this is open to scientific investigation. Thus, inevitably, the developments I have tried to sketch will affect Man's attitude to himself.

References

Beach, F. A. (1955) The descent of instinct. *Psychol. Rev.* **62**, 401–10.

Berkun, M. M., Kessen, M. L., and Miller, N. E. (1952) Hunger-reducing effects of food by stomach fistula versus food by mouth measured by a consummatory response. *J. comp. physiol. Psychol.* **45**, 550–4.

Blest, A. D. (1957) The evolution of protective displays in the Saturnioidea and Sphingidae (Lepidoptera). *Behaviour* **11**, 257–309.

Bowlby, J. (1960) Ethology and the development of object relations. *Int. J. Psycho-Analysis* **41**, 313–17.

Cain, A. J. (1964) The perfection of animals. *Viewpoints Biol.* **3**, 37–63.

Craig, W. (1918) Appetites and aversions as constituents of instincts. *Biol. Bull.* **34**, 91–107.

Crossley, S. A. (1963) Doctor's thesis, Oxford; see also Tinbergen 1965*b*.

Cullen, E. (1957). Adaptations to cliff-nesting in the Kittiwake. *Ibis* **99**, 275–302

—— (1961) The effect of isolation from the father on the behaviour of male Three-spined Sticklebacks to models. *Tech.* (*Final*) *Rep. on Contract AF* 61(052)–29, *USAFRDC.*

Darwin, C. (1859) *On the origin of species by natural selection.* London.

—— (1872) *The expression of emotions in man and animals.* London.

Desmedt, J. E. and Monaco, P. (1961). Mode of action of the efferent olivo-cochlear bundle on the inner ear. *Nature, Lond.* **192**, 1263–5.

Dethier, V. G. and Bodenstein, D. (1958) Hunger in the Blowfly. *Z. Tierpsychol.* **15**, 129–40.

Eccles, J. C. (1953) *The neurophysiological basis of mind.* Oxford University Press.

Eibl-Eibesfeldt, I. (1963) Angeborenes und Erworbenes im Verhalten einger Säuger. *Z. Tierpsychol.* **20**, 705–54.

Grohmann, J. (1939) Modifikation oder Funktionsreifung? *Z. Tierpsychol.* **3**, 132–44.

Harris, G. W., Michael, R. P., and Scott, P. P. (1958) Neurological site of action of stilboestrol in eliciting sexual behaviour. *Ciba Foundation Symposium on the Neurological Basis of Behaviour*, London, pp. 236–51.

Hassenstein, B. (1961) Wie sehen Insekten Bewegungen? *Naturwissenschaften* **48**, 207–14.

Heinroth, O. (1911) Beiträge zur Biologie, namentlich Ethologie und Psychologie der anatiden. *Verh. Ver. Int. orn. Congr.*, pp. 589–702.

—— and Heinroth, M. (1928) *Die Vögel Mitteleuropas.* Berlin.

Hinde, R. A. (1965) The integration of the reproductive behaviour of female canaries. *Sex and behaviour* (edited by F. A. Beach), pp. 381–416. New York.

—— (1966) *Animal behaviour.* New York.

Holst, E. von (1939) Entwurf eines Systems der lokomotorischen Periodenbildungen bei Fischen. *Z. vergl. Physiol.* **26**, 481–528.

—— and Mittelstaedt, H. (1950) Das Reafferenzprinzip. *Naturwissenschaften* **37**, 464–76.

—— and St. Paul, U. von (1960) Vom Wirkungsgefüge der Triebe. *Naturwissenschaften* **47**, 409–22.

Huxley, J. S. (1914) The courtship habits of the Great Crested Grebe (*Prodiceps cristatus*); with an addition to the theory of sexual selection. *Proc. zool. Soc. Lond.* 419–562.

—— (1923) Courtship activities in the Red-throated Diver (*Colymbus stellatus* Pontopp.); together with a discussion on the evolution of courtship in birds. *J. Linn. Soc.* **35**, 253–92 (1923).

—— (1963) Lorenzian ethology. *Z. Tierpsychol.* **20**, 402–09.

Jennings, H. S. (1923) *The behavior of lower organisms.* New York.

Knoll, M. D. (1956) Ueber die Entwicklung einiger Funktionen im Auge des Grasfrosches. *Z. vergl. Physiol.* **38**, 219–37.

Konishi, M. (1965) The role of auditory feedback in the control of vocalisation in the White-crowned Sparrow. *Z. Tierpsychol.* **22**, 770–83.

Kuo, Z. Y. (1932) Ontogeny of embryonic behaviour in Aves. IV. *J. comp. Psychol.* **14**, 109–21.

Lack, D. (1947) *Darwin's finches.* London.

Lehrman, D. S. (1953) A critique of Konrad Lorenz's theory of instinctive behaviour. *Q. Rev. Biol.* **28**, 337–63.

—— (1961) Gonadal hormones and parental behaviour in birds and infrahuman mammals. *Sex and internal secretions* (edited by W. C. Young). Baltimore.

Lissmann, H. W. (1946) The neurological basis of the locomotory rhythm in the spinal Dogfish (*Scyllium canicula, Acanthias vulgaris*) II, *J. exp. Biol.* **23**, 162–76.

Lorenz, K. (1927) Beobachtungen an Dohlen. *J. Orn., Lpz.* **75**, 511–19.

—— (1931) Beiträge zur Ethologie sozialer Corviden. *J. Orn., Lpz.* **79**, 67–120.

—— (1935). Der Kumpan in der Umwelt des Vogels. *J. Orn., Lpz.* **83**, 137–213; 289–413

—— (1937) Ueber die Bildung des Instinktbegriffs. *Naturwissenschaften* **25**, 289–300; 307–18; 324–31.

—— (1939) Vergleichende Verhaltensforschung. *Zool. Anz.* Suppl. **12**, 69–102.

—— (1969) *Evolution and modification of behavior.* Chicago.

Morgan, L. (1896) *Habit and instinct.* London.

Pavlov, J. P. (1926) *Die höchste Nerventätigkeit (Das Verhalten) von Tieren.* München.

Perdeck, A. C. An experiment on the ending of autumn migration in Starlings. *Arde*, **52**, 133–40.

Roeder, K. C. (1962) Neural mechanisms of animal behavior. *Am. Zool.* **2**, 105–15.

Schneirla, T. C. (1966) Behavioral development and comparative psychology. *Q. Rev. Biol.* **41**, 283–302.

Schneirla, T. C., Rosenblatt, J. S., and Tobach, E. (1963) Maternal behavior in the Cat. *Maternal behavior in mammals* (edited by H. Rheingold). New York.

Schutz, F. (1965) Sexuelle Prägung bei Anatiden. *Z. Tierpsychol.* **22**, 50–103.

Sevenster-Bol, A. C. A. (1962) On the causation of drive-reduction after a consummatory act. *Archs néerl. Zool.* **15**, 175–236.

Sherrington, C. S. (1906) *Integrative action of the nervous system.* London.

Sperry, R. W. (1959) The growth of nerve circuits. *Scient. Am.* **201**, 68–76.

Sutherland, N. S. (1962) The methods and findings of experiments on the visual discrimination of shape by animals. *Exp. psychol. Soc. Monogr.* 1.

Thomas, E. and Schaller, F. (1954) Das Spiel der optisch isolierten, jungen Kaspar-Hauser-Katze. *Naturwissenschaften* **41**, 557–8.

Thorpe, W. H. (1961) *Bird song.* London.

—— (1963) *Learning and instinct in animals.* London.

Tinbergen, N. (1953) *Social behaviour in animals.* London.

—— (1964) Aggression and fear in the normal sexual behaviour of some animals. *The pathology and treatment of sexual deviation* (edited by J. Rosen), pp. 3–23. Oxford University Press.

—— (1965*a*) Some recent studies of the evolution of sexual behaviour. *Sex and behavior* (edited by F. A. Beach), pp. 1–34.

—— (1965*b*) Behavior and natural selection. *Ideas in modern biology* (edited by J. A. Moore), pp. 521–42. New York.

—— (1962) *et al.* Egg-shell removal by the Black-headed Gull *Larus ridibundus* L.; a behavioural component of camouflage. *Behaviour* **19**, 74–118.

—— and Kuenen, D. J. (1939) Ueber die auslösenden und die richtunggebenden Reizsituationen der Sperrbewegung von jungen Drosseln. *Z. Tierpsychol.* **3**, 37–60.

Verwey, J. (1930) Die Paarungsbiologie des Fischreihers. *Zool. Jahrb. Allg. Zool. Physiol.* **48**, 1–120.

Waterman, T. H., Wiersma, G. A. G., and Bush, B. M. H. (1964) Afferent visual responses in the optic nerve of the crab, *Podophthalmus. J. Comp. Cell. Physiol.* **63**, 133–55.

Wickler, W. (1962) Zur Stammesgeschichte funktionell korrelierter Organ- und Verhaltensmerkmale. *Z. Tierpsychol.* **19**, 29–64.

Whitman, C. O. (1919) *The behavior of pigeons. Carnegie Institute, Washington, Publication* 257, Vol. 3, pp. 1–161.

Wiepkema, P. R. (1961) An ethological analysis of the reproductive behaviour of the Bitterling. *Archs néerl. Zool.* **14**, 103–99.

Author Index

Subject Index